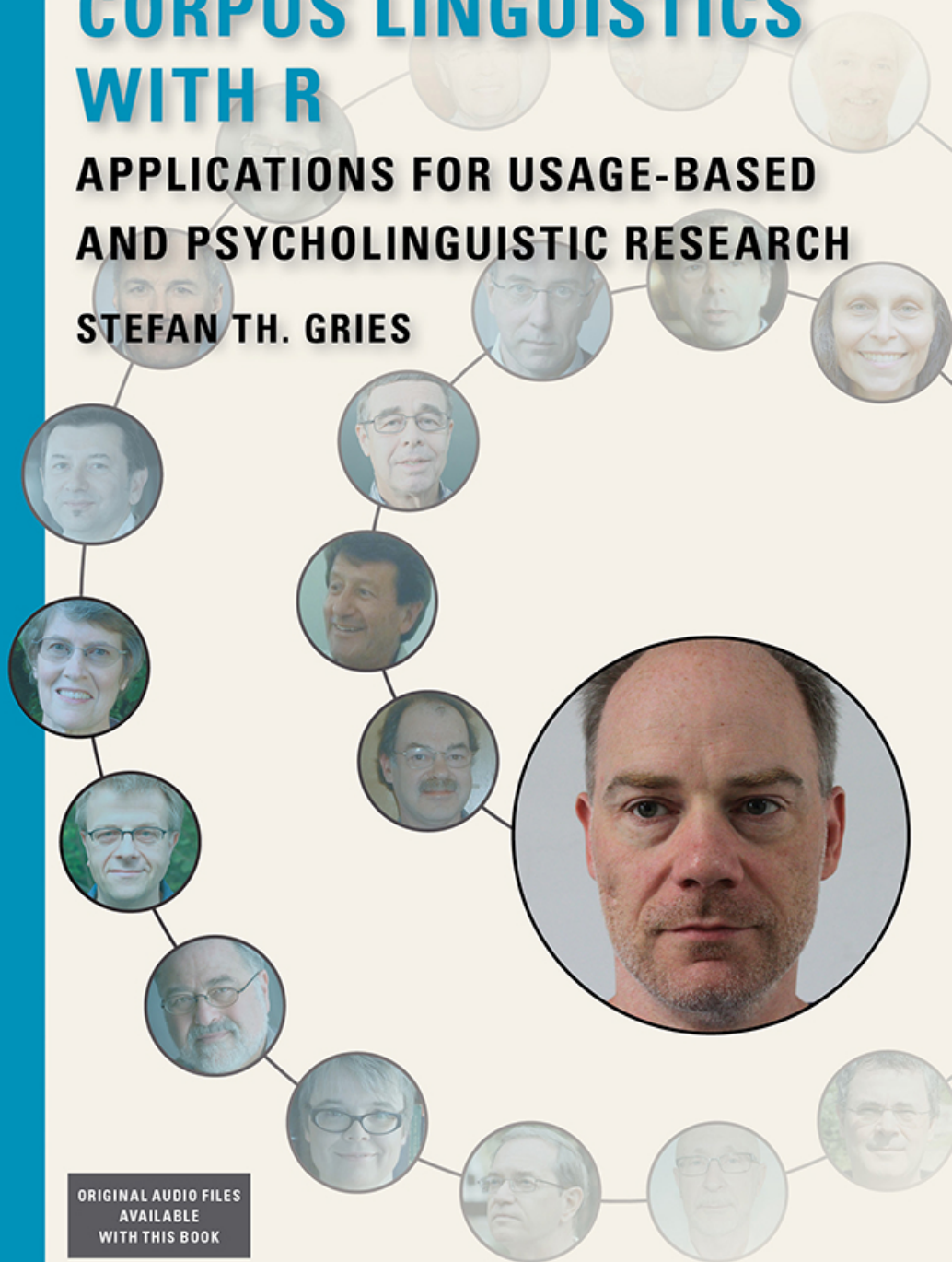


# TEN LECTURES ON CORPUS LINGUISTICS WITH R

APPLICATIONS FOR USAGE-BASED  
AND PSYCHOLINGUISTIC RESEARCH

STEFAN TH. GRIES



ORIGINAL AUDIO FILES  
AVAILABLE  
WITH THIS BOOK

DISTINGUISHED LECTURES IN COGNITIVE LINGUISTICS

BRILL

## Ten Lectures on Corpus Linguistics with R

# Distinguished Lectures in Cognitive Linguistics

*Edited by*

Fuyin (Thomas) Li (*Beihang University, Beijing*)

*Guest Editors*

Jing Du, Na Liu and Shan Zuo (*Beihang University*)

*Editorial Assistants*

Jing Du, Na Liu and Cuiying Zhang (*doctoral students at  
Beihang University*)

*Editorial Board*

Jürgen Bohnemeyer (*State University of New York at Buffalo*) – Alan Cienki  
(*Vrije Universiteit (VU) at Amsterdam, Netherlands and Moscow State  
Linguistic University, Russia*) – William Croft (*University of New Mexico,  
USA*) – Ewa Dąbrowska (*Northumbria University, UK*) – Gilles Fauconnier  
(*University of California at San Diego, USA*) – Dirk Geeraerts (*University of  
Leuven, Belgium*) – Nikolas Gisborne (*The University of Edinburgh, UK*) –  
Cliff Goddard (*Griffith University, Australia*) – Stefan Th. Gries (*University of  
California at Santa Barbara and Justus-Liebig-Universität Giessen, USA*) –  
Laura A. Janda (*University of Tromsø, Norway*) – Zoltán Kövecses (*Eötvös  
Loránd University, Hungary*) – George Lakoff (*University of California at  
Berkeley, USA*) – Ronald W. Langacker (*University of California at San Diego,  
USA*) – Chris Sinha (*Hunan University, China*) – Leonard Talmy (*State  
University of New York at Buffalo, USA*) – John R. Taylor (*University of Otago,  
New Zealand*) – Mark Turner (*Case Western Reserve University, USA*) –  
Sherman Wilcox (*University of New Mexico, USA*) – Phillip Wolff (*Emory  
University, USA*) Jeffrey M. Zacks (*Washington University in St. Louis, USA*)

*Distinguished Lectures in Cognitive Linguistics* publishes the keynote lectures series given by prominent international scholars at the China International Forum on Cognitive Linguistics since 2004. Each volume contains the transcripts of 10 lectures under one theme given by an acknowledged expert on a subject and readers have access to the audio recordings of the lectures through links in the e-book and QR codes in the printed volume. This series provides a unique course on the broad subject of Cognitive Linguistics. Speakers include George Lakoff, Ronald Langacker, Leonard Talmy, Laura Janda, Dirk Geeraerts, Ewa Dąbrowska and many others.

The titles published in this series are listed at [brill.com/dlcl](http://brill.com/dlcl)

# Ten Lectures on Corpus Linguistics with R

*Applications for Usage-Based  
and Psycholinguistic Research*

*By*

Stefan Th. Gries



BRILL

LEIDEN | BOSTON



## Library of Congress Cataloging-in-Publication Data

Names: Gries, Stefan Thomas, 1970- author.

Title: Ten lectures on corpus linguistics with R : applications for usage-based and psycholinguistic research / by Stefan Th. Gries.

Description: Leiden ; Boston : Brill, 2020. | Series: Distinguished lectures in cognitive linguistics, 2468-4872 ; vol. 23 | Includes bibliographical references.

Identifiers: LCCN 2019040920 (print) | LCCN 2019040921 (ebook) | ISBN 9789004410336 (hardback) | ISBN 9789004410343 (ebook)

Subjects: LCSH: Corpora (Linguistics) | Linguistic analysis (Linguistics) | Psycholinguistics.

Classification: LCC P128.C68 G754 2020 (print) | LCC P128.C68 (ebook) | DDC 410.1/88—dc23

LC record available at <https://lcn.loc.gov/2019040920>

LC ebook record available at <https://lcn.loc.gov/2019040921>

Typeface for the Latin, Greek, and Cyrillic scripts: “Brill”. See and download: [brill.com/brill-typeface](http://brill.com/brill-typeface).

ISSN 2468-4872

ISBN 978-90-04-41033-6 (hardback)

ISBN 978-90-04-41034-3 (e-book)

Copyright 2020 by Stefan Th. Gries. Reproduced with kind permission from the author by Koninklijke Brill NV, Leiden, The Netherlands.

Koninklijke Brill NV incorporates the imprints Brill, Brill Hes & De Graaf, Brill Nijhoff, Brill Rodopi, Brill Sense, Hotei Publishing, mentis Verlag, Verlag Ferdinand Schöningh and Wilhelm Fink Verlag.

All rights reserved. No part of this publication may be reproduced, translated, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior written permission from the publisher.

Authorization to photocopy items for internal or personal use is granted by Koninklijke Brill NV provided that the appropriate fees are paid directly to The Copyright Clearance Center, 222 Rosewood Drive, Suite 910, Danvers, MA 01923, USA. Fees are subject to change.

This book is printed on acid-free paper and produced in a sustainable manner.

# Contents

	Note on Supplementary Material	VII
	Preface by the Series Editor	VIII
	Preface by the Author	X
	About the Author	XI
1	Corpus Linguistics: the (Methods of the) Field and Its Relation to Cognitive Linguistics	1
2	On—and/or Against—Frequencies	38
3	Frequency: Practice with R	72
4	On Recency and Dispersion	90
5	Dispersion: Practice with R	131
6	On Association	147
7	Association: Practice with R	182
8	On Context	199
9	Concordance, Surprisal, Entropy: Practice with R	234
10	Corpus-Linguistic Applications in Cognitive/Usage-Based Explorations of Learner Language	249
	References	285
	About the Series Editor	294
	Websites for Cognitive Linguistics and C1FL Speakers	295



## Note on Supplementary Material

All original audio-recordings and other supplementary material, such as hand-outs and PowerPoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via a QR code for the print version of this book. In the e-book both the QR code and dynamic links will be available which can be accessed by a mouse-click.

The material can be accessed on [figshare.com](http://figshare.com) through a PC internet browser or via mobile devices such as a smartphone or tablet. To listen to the audio recording on hand-held devices, the QR code that appears at the beginning of each chapter should be scanned with a smart phone or tablet. A QR reader/scanner and audio player should be installed on these devices. Alternatively, for the e-book version, one can simply click on the QR code provided to be redirected to the appropriate website.

This book has been made with the intent that the book and the audio are both available and usable as separate entities. Both are complemented by the availability of the actual files of the presentations and material provided as hand-outs at the time these lectures were given. All rights and permission remain with the authors of the respective works, the audio-recording and supplementary material are made available in Open Access via a CC-BY-NC license and are reproduced with kind permission from the authors. The recordings are courtesy of the China International Forum on Cognitive Linguistics (<http://cifcl.buaa.edu.cn/>), funded by the Beihang University Grant for International Outstanding Scholars.



The complete collection of lectures by Stefan Th. Gries can be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.c.4617035>

## Preface by the Series Editor

The present text, entitled *Ten Lectures on Corpus Linguistics with R for Usage-based and Psycholinguistic Research* by Stefan Th. Gries, is a transcribed version of the lectures given by Professor Gries in October 2018 as the forum speaker for the 18th China International Forum on Cognitive Linguistics.

*The China International Forum on Cognitive Linguistics* (<http://cifcl.buaa.edu.cn/>) provides a forum for eminent international scholars to give lectures on their original contributions to the field of cognitive linguistics. It is a continuing program organized by several prestigious universities in Beijing. The following is a list of organizers for CIFCL 18.

Organizer:

Fuyin (Thomas) Li: PhD/Professor, Beihang University

Co-organizers:

Yihong Gao: PhD/Professor, Peking University

Baohui Shi: PhD/Professor, Beijing Forestry University

Yuan Gao: PhD/Professor, University of Chinese Academy of Sciences

Sai Ma: PhD, Capital Normal University

The text is published, accompanied by its audio disc counterpart, as one of the *Distinguished Lectures in Cognitive Linguistics*. The transcriptions of the video, proofreading of the text and publication of the work in its present book form have involved many people's strenuous efforts. The initial transcripts were completed by Shan Zuo, Jinmei Li, Hongxia Jia, Chenxi Niu, Shu Qi, Mengxue Duan, Junjie Lu and Na Liu. Na Liu and Shan Zuo made revisions to the whole text. We editors then made word-by-word and line-by-line revisions. To improve the readability of the text, we have deleted the false starts, repetitions, fillers like *now, so, you know, OK, and so on, again, of course, if you like, sort of*, etc. Occasionally, the written version needs an additional word to be clear, a word that was not actually spoken in the lecture. We have added such words within single brackets [...]. To make the written version readable, even without watching the film, we've added a few "stage directions", in italics also within single brackets: [...]. These describes what the speaker was doing, such

as pointing at a slide, showing an object, etc. Professor Gries made final revisions to the transcriptions; the published version is the final version approved by the speaker.

*Thomas Fuyin Li*

Beihang University

*thomasli@buaa.edu.cn*

*Jing Du*

Beihang University

*millydu1019@buaa.edu.cn*

## Preface by the Author

I was very honored and happy to be invited back to Beijing for another round of talks at the China International Forum on Cognitive Linguistics. Like last time, it was a wonderful opportunity to discuss work that I care about a lot (by others and myself) in a context that allowed me to present a much bigger picture, discuss much more of contents that are—due to space constraints—usually shortened or even omitted, and include the practical corpus-linguistic component that I think has done so much for Cognitive Linguistics, and I am grateful to have been offered this chance again.

I am also very grateful to Prof. Fuyin (Thomas) Li for his hospitality and energy in making this happen and ensuring that everything went just as well as last time; my thanks of course also go to all members of the organizing team, the student team taking care of all logistical aspects of my stay, and the editorial team responsible for making this publication happen, Thomas Li, Jing Du, Na Liu, and Shan Zuo—if this book turns out to be informative to the cognitive-linguistic community, credit is due to their diligent work as well.

*Stefan Th. Gries*

UC Santa Barbara & JLU Giessen

January 2019

## About the Author

Stefan Th. Gries earned his M.A. and Ph.D. degrees at the University of Hamburg in 1998 and 2000 and is currently (Full) Professor of Linguistics in the Department of Linguistics at the University of California, Santa Barbara (UCSB) as well as a part-time (Full/W3-) Professor of Corpus Linguistics (Quantitative Methods) of the Justus-Liebig-Universität Giessen.

Methodologically, Gries is a quantitative corpus linguist at the intersection of corpus linguistics, cognitive linguistics, and computational linguistics, who uses statistical methods to investigate linguistic phenomena (corpus-linguistically and experimentally) and tests and develops corpus-linguistic and statistical methods. Theoretically, he is a cognitively-oriented usage-based linguist. Gries has authored three books, co-edited seven volumes, and has (co-)authored dozens of articles, many in the leading peer-reviewed journals of his fields (*Cognitive Linguistics* and *International Journal of Corpus Linguistics*). He is general editor of the international peer-reviewed journal *Corpus Linguistics and Linguistic Theory*, co-editor-in-chief of the *Journal of Research Design and Statistics in Linguistics and Communication*, associate editor of *Cognitive Linguistic Studies*, and performs editorial functions for about a dozen other international peer-reviewed journals and book series.





## Corpus Linguistics: the (Methods of the) Field and Its Relation to Cognitive Linguistics

Thank you very much for the introduction and of course also for the invitation to be here. As Professor Li said, it is going to be ten talks on corpus linguistics. It's basically what this title says here and its relationship to cognitive linguistics. There will be a lot of corpus linguistics stuff, and there will be lot of quantitative and statistical stuff but I hope to be able to show that a variety of things that have been going on in cognitive or usage-based linguistics can benefit from a certain kind of improvement or a certain set of improvements. Basically, over the talks, there will always be a coupling of a more theoretical talk and more empirical or practical talk that will then culminate hopefully at the end with one talk that shows how many of these things got put together in a variety of different case studies.

I want to start at a relatively slow pace to give those of you who haven't worked with corpora at all yet a little bit of a legs up to see what is corpus linguistics all about. And then in the remainder of this talk, I want to talk a little bit about how some of the corpus stuff is seen in cognitive linguistics, usage-based linguistics, and to at least some extent, I want to talk about how I think some of these views are mistaken and maybe should be revised a little bit. So, at the beginning, I'll talk about corpus linguistics in general, and then move on to cognitive stuff per se.

So very basic then as the beginning, what is corpus linguistics? I've talked a little bit about this actually a few years ago when I was here, because there's been a lot of debate on whether it's a theory or a method. Some people consider it a completely theoretical approach on a par with cognitive linguistics, usage-based linguistics, or other things like that whereas other people as the ones that are cited here [McEnery & Wilson 1996, Meyer 2002, Bowker & Pearson 2002, Hardie 2008] basically consider it more of a method.



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611081>

## what is corpus linguistics?

- There has been a lot of debate about whether it's
  - a **theory** (Leech 1992, Stubbs 1993, Tognini-Bonelli 2001, Teubert 2005) or
  - a **method** (McEnery & Wilson 1996, Meyer 2002, Bowker & Pearson 2002, Hardie 2008)
- just like when I was here 4 years ago, I still consider it a method
- but of course it's more like a family of methods, which all have some things in common
  - they are applied to a certain kind of data source - **corpora** (duh)
  - as such they are all based on the **presence or absence of character strings**

FIGURE 1

Just like a number of years ago, I still consider it a method as opposed to a theory in particular, because we would never call any other kind of method in linguistics a theory. There's no eye-movement linguistics, and that's called a theory or something like that or eye-tracking linguistics. But as you will see, it's probably more useful to actually consider it a family of methods, because they all have something in common. Some of these things are actually not yet fully appreciated, I think, in a lot of ways, in which usage-based linguistics is pursuing.

One obvious thing is that corpus linguistics relies on corpora. That's not exactly sensational. The other thing is that everything we do with corpora is based on the presence or absence of character strings in a sense. That's something I do want to talk a little bit more about because it has implications on how we as cognitive or usage-based linguists try to work with corpora.

Here's kind of a schema of the different kinds of data one might look at in linguistics. They're kind of ordered a little bit in terms of naturalness of what you're doing or what you're looking at.

At the very top, we have corpora with written texts, newspapers or blogs. I'll define corpus in a little bit more detail in a moment, but [[they are]] essentially ideally large collection of texts that contain newspaper language, blogs, or something like that. As we go down this cline, you can see that we're becoming more and more specialized and more and more different from the ways that we usually and naturally interact with language. For instance, somewhere here

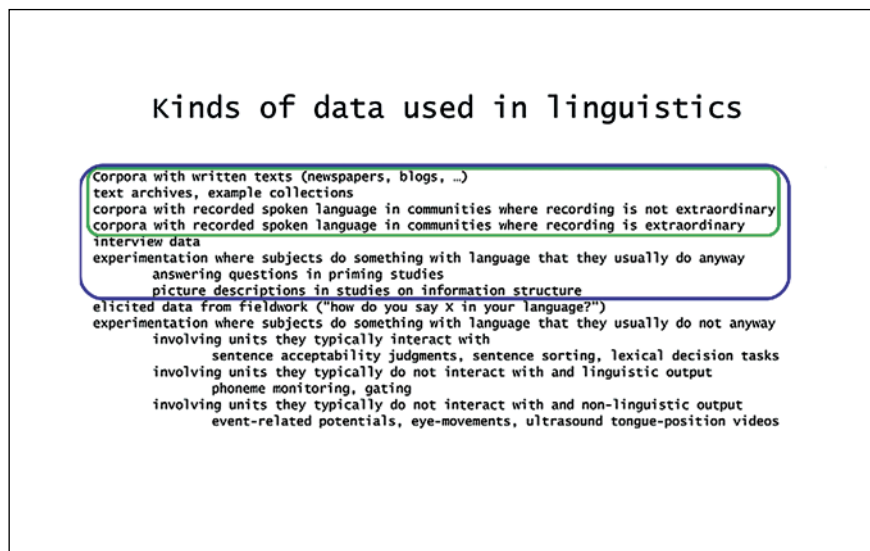


FIGURE 2

in the middle, we have recorded spoken language with interview data, which are not necessarily the most natural way in which people use language such as in conversations.

Then below that we have, for instance, experimentation where subjects do something with language that they usually do anyway. For instance, if you do a priming experiment where subjects don't actually know exactly what the experiment is about, but they answer questions or they describe a picture, then obviously answering questions and describing pictures is something we do all the time anyway. So, the context here is experimental, but it's still relatively natural.

Then here at the bottom, for instance, we have things that are quite remote from what we normally do. We have experimentation where subjects do something with language we don't usually do. Usually you don't sit down to have event-related potentials measured while you're having your tongue-position measured with ultrasound. So the critical thing here is that depending on how you want to define corpora or corpus linguistics, you would either take the upper part now that is highlighted in blue [referring to the part above the "elicited data"], meaning you would consider something like picture descriptions provided in an experiment or questions answered in a priming study, you would consider those still part of corpus data, or you might adopt a kind of narrower view when you say only these parts here [pointing to the contents circled in green] constitute corpora in the narrower sense of the term.

## what is a prototypical corpus?

- A prototypical corpus is
  - a machine-readable collection of texts (spoken or written)
  - that were produced in a natural communicative setting
  - representative and balanced with respect to a particular variety/register/genre
  - that are compiled to be analyzed linguistically

FIGURE 3

What they lead to us essentially is that the corpus is actually just the prototype category, which of course, is something that as cognitive linguists we are all too familiar with. A prototypical corpus would be a machine-readable collection of texts, typically, produced in a natural communicative setting, so that's where experiments might constitute a borderline case, and then they would be representative and balanced with respect to a particular variety or register, or genre or something like that. Finally, the idea is they have been compiled to be analyzed linguistically.

To get a better understanding of what each of these things means, let's unpack them a little bit. Some of those are pretty straightforward, others maybe not so much.

[[As for]] machine readable, I think it's pretty straightforward. Today you will find that pretty much all of the corpora that you would get access to would be plain text files, typically, hopefully these days Unicode so that all sorts of writing systems can be accommodated. There's a range of formats that you will typically find corpora in: SGML was, for a long time, one of the most dominating standards. By now, I guess it's XML annotation that you would find most frequently and those of course help in order to work with any kind of annotation that you have in your corpora in a nicely replicable way. Hopefully, you will not find any corpora anymore as \*.doc files, as a Word file, or something like that or actually on paper. Some corpora in fact come with relatively sophisticated retrieval software. So a corpus from which you will see a lot of data mentioned

## what is a prototypical corpus?

- "Machine-readable"
  - virtually all corpora are stored in the form of plain text files (ASCII or *Unicode*, typically/hopefully UTF-8) that can be loaded, manipulated, and processed platform-independently
  - frequent formats of annotated corpora
    - SGML
    - XML
  - what you will not find much (anymore, hopefully, please!)
    - corpus files as \*.doc
    - corpus data on paper
  - some corpora come with sophisticated retrieval software (e.g., ICE-GB)

FIGURE 4

later is this one here, The International Corpus of English, the British component (ICE-GB). For those of you who don't know it, that is a one-million-word corpus, representative roughly of British English of the 1990s. The nice thing about it is first, in terms of percentage, it has a very large spoken component, namely 60% of the corpus is actually spoken data, and 'only' 40% are written data. Secondly, the corpus is tagged and parsed and manually corrected. So the parse trees are pretty reliable, compared to some other kinds of automatic parses that you will find in the relevant literature. So much for machine readable in this context.

Second, [[a prototypical corpus is]] produced in a natural communicative setting. What do I mean by this is? The texts were written or spoken, or theoretically signed or theoretically other measurements that were taken in a natural communicative setting. They were written and produced for some authentic communicative purpose as opposed to, let's collect those things for a corpus, which is sort of the middle part of the beginning continuum of different kinds of corpora, that's where this criterion would distinguish one set of corpus data from the other. For instance, journalese, newspaper language in corpora, would obviously meet this criterion, because journalists write articles to communicate something to the newspapers, and not because they know that that kind of stuff will later end up in a corpus, obviously. At the same time, to use a spoken example, if you record someone's speech for a week, then hopefully you will get some authentic discourse from it, even though, these

## What is a prototypical corpus?

- "produced in a natural communicative setting"
- the texts were spoken or written or ... for some authentic communicative purpose, not for putting them into a corpus
- example: journalese in corpora meets this criterion
  - journalists write articles to communicate something in their newspapers, not to fill a linguist's corpus
- example: if I record someone's speech for a week, I will hopefully obtain authentic discourse (even though all interlocutors should know they are being recorded)

FIGURE 5

## What is a prototypical corpus?

- "representative with respect to a particular variety"
- the different parts of the variety I am interested in are all manifested in the corpus
- example: phonological reduction in the speech of Californian adolescents
  - if I only record Californian adolescents in their peer groups, I would fail to collect data on a whole variety of additional groups of interlocutors
    - parents
    - teachers
    - ...

FIGURE 6

days, all interlocutors should know that they are being recorded. Typically, whatever kind of awkwardness this might induce at the beginning that goes away pretty quickly.

Then, “representative with respect to a particular variety”. We’re now beginning to look at some critical or difficult issues. First, what that means is that different parts of the variety that you’re interested in are all manifested in the corpus. A corpus is a sample of language, typically that is supposed to represent a larger population that we cannot collect altogether. That larger population of language will have many different parts, many different registers, or genres, or something like that, or varieties and the idea is that a corpus is representative to the extent that all those are covered. An example I’ve used before: if you look at phonological reduction in a spoken corpus of Californian adolescents, then if you only record them while they are talking to members of their peer groups, then you miss out on a whole bunch of other kinds of registers, namely, what happens when they talk to parents, what happens when they talk to teachers, and so on. So a truly representative corpus that is supposed to cover that kind of population would include at least some sample recordings from all of these different kinds of subgroups.

### what is a prototypical corpus?

- “balanced with respect to a particular variety”
  - not only should all parts of the variety I am interested in be included
  - also, the proportions of the parts with which they are represented in the sample (i.e., the corpus) should reflect the proportions with which they occur in the population
  - example: if dialogs make up 65% of the speech of Californian adolescents, 65% of my corpus of the speech of Californian adolescents should be dialog data

FIGURE 7



That, together then with the next one, “balanced with respect to a particular variety”, is the second component of those two things that are mostly a theoretical ideal. “Balanced” means that you shouldn’t only have all the parts of the variety that you’re interested in, which is what representativeness was about. It also means that the sizes of the corpus parts ideally would reflect the proportions of the registers or varieties in the population. If we actually had any reliable data on how much on average does a person use language orally, and how much does a person do that in writing, then ideally a corpus that we would have would reflect those proportions in its sampling scheme. If dialogs make up on average 65% of the speech of Californian adolescents, then if you compile a corpus of that variety, then 65% of that should be dialog data.

Now, the problem of course with this is we can only look at a small sample of Californian adolescents, and the percentage will vary. There is going to be a huge degree of inter-speaker variability that will make any kind of simple average be rather unreliable. Second, it’s not obvious how we would measure these proportions, I mean, how much in terms of minutes or words or sentences, what is the unit there that goes into your computation of percentages? Also, how would you measure the importance of any particular variety? In cognitive linguistics, in many areas in discourse and functional linguistics, conversational speech is considered primary: It’s supposed to be the most basic use of language, the one that everyone engages in. Obviously, there are a lot of languages

### what is a prototypical corpus?

- “balanced with respect to a particular variety”
- problems
  - we can only measure a small sample of Californian adolescents: the percentage will vary
  - how would we measure the proportions: in terms of minutes, words, sentences, ...?
  - how would we measure the importance of any linguistic variety?
    - usually, conversational speech is considered primary ...
    - ... but a single newspaper headline may have a more radical effect on any speaker’s linguistic system than many hours of conversational speech
- balancedness = theoretical ideal

FIGURE 8

out there for which there isn't even a written alphabet or something like that so obviously, conversational speech might be primary there.

On the other hand, we all know examples from cases where non-conversational speech has had a much higher impact on what speakers do, on how speaker's linguistic system changes than anything else. One single, witty or particularly interesting newspaper headline, or line in a movie or something like that, might immediately be in everyone's consciousness when triggered. It's not always clear that conversational speech is always primary with regard to everything that might happen in a language. So balancedness, for those reasons, is essentially a theoretical ideal when we usually don't really know what these percentages are, and so our best guess is always 'sample broadly' in the hope that would come close to a representative and balanced sample.

Now, with regard to these criteria, and if corpora are a prototype category, then what would be more marginal corpora? One example would be a collection of a second language learner or foreign language learner essays, as they are now more and more used in a second language acquisition research or learner corpus research. This is more marginal because, such essays as they enter into corpora are usually not produced in a natural communicative setting. I mean teachers assign topics, it's not like you can always write what you want to write on, they might impose time or word limits, and they grade, which leads to avoidance strategies, so students don't use language as freely as they would if

### what is a more marginal corpus?

- A collection of L2/FL learner essays
  - such essays are usually not produced in a natural communicative setting: teachers
    - assign topics, impose time/word limits, grade, ...
- a database of texts
  - which was usually not compiled for linguistic analysis
  - which was not intended to be representative or balanced of any particular linguistic variety
  - example: a publisher of some popular computing periodical makes all issues of the previous year available on a website
- dialect corpora of people from different dialects
- reading sentences
- the DCIEM Map Task Corpus
- eye-tracking corpora ...

FIGURE 9

they know things don't get graded and so on. All of these things of course eat into the natural communicative setting kind of criterion.

A second example would be a database of texts not compiled for linguistic analysis, and therefore not intended to be representative or balanced of anything. If you look at a computing periodical, some website online that discusses computer stuff like cnet.com or zdnet.com or something like that—if all the issues of a certain year I made available on the website, then that's more like a database than a corpus. That's been produced in a natural communicative settings, but at the same time, it's not balanced or representative with regard to anything.

Yet another example of a more marginal corpus typically at least would be something like this, so dialect corpora, for instance, we have people from different dialects read sentences and the idea is to track dialectal variation or sociolinguistic variation in how they read these sentences. The corpus is more marginal, because obviously, the communicative setting is not particularly natural. If you go to someone and then say “read out these sentences one by one”, that's not a most natural thing to do and so the corpus is much more unrepresentative of what normally would happen.

Then there are way more experimental types of corpora. The DCIEM Map Task Corpus is a corpus of a speech collected by people who were sleep-deprived and under drugs. That was maybe unsurprisingly an army project. The idea was if soldiers are in the field and they are sleep-deprived and they're under drugs, and they have to describe a route on a map, then how much do these circumstances influence the clarity of the description? I hope we agree that that is not exactly representative of the normal kind of speech that we would use.

Then there's even more exotic—note the air quotes—things like eye-tracking corpora where the idea is that you don't even record spoken or written language, but what you're looking at is eye movements as people read texts. Those kinds of things can be really extremely interesting because the idea would be that the eye movements, for instance, reflect areas of syntactic processing difficulty: If people move back from what's supposed to be the end of a sentence, they backtrack a few words to resolve a garden path or something like that, then obviously, in an eye-tracking corpus, you would have a trace of these movements, and you could try and correlate the amount of time spent on particular words and the amount of backtracking to certain phrases with the processing complexity associated with those linguistic elements.

Now what would be even more marginal corpora? For instance, example collections of words or sentences. For instance, I am still, after all those years, compiling a collection of blends, so expressions like *brunch* and *motel* that I

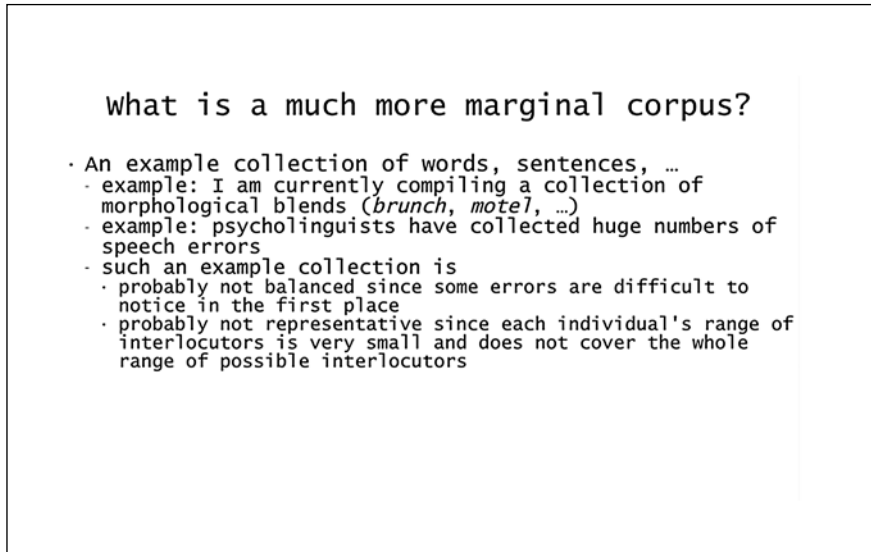


FIGURE 10

talked about here a few years ago. What I do there basically is whenever I see one, I might write it down. That's obviously not particularly representative, because there are a lot of factors that might lead to me missing a certain blend. Some things I will recognize very easily, so they have a higher chance of ending up in the corpus; some things I might not recognize as a blend very quickly or not at all, so I don't write it down, even though it is a good example of a blend. That basically means that the corpus collection process, so to speak, here will be tainted by ease of perception or ease of identifying something and it's not going to be representative of anything other than my own perception, probabilities or difficulties. The same thing [[is the case]] with speech error data collections. In the 1970s, there was a lot of work where people wrote down every speech error that occurred within earshot, and they noted down what they heard and they noted down what the intended target was to then later analyze how does speakers produce, how did it differ from what they actually intended to do? The same thing [[is happening]] here: Some errors will be very difficult to notice in the first place, so their chance of making it into a corpus or a database are much less than for something which is very easy to hear. A blend like *absotively*, out of *absolutely* and *positively*, you cannot not hear that, it's super obvious you would never miss that so writing that down, hearing that thing, it making it into a corpus is very obvious. If your speech error consists of devoicing of a word final consonant, someone might just not hear that. It happened, but it didn't make it into the corpus because it was harder to hear. These

### what does "presence or absence of character strings" mean?

- Corpora do not really/directly provide what most (cognitive) linguists are interested in
    - meaning
    - communicative function/intention
    - information structure
    - cognition/processing
    - language proficiency/dialect, ...
  - they provide information on the presence or absence of character strings
    - in (certain parts/locations of) corpora
    - in the presence or absence of other character strings
- note: character strings can be anything: morphemes, words, constructions, anything including and especially markup/annotation!

FIGURE 11

kinds of databases/corpora then would be less great examples of what people might call a corpus, also because of the representativity problem indicated here. If you record speech errors with a bunch of interlocutors as you interact with people throughout the day, that's probably not particularly representative, because we all have our pockets and preferred types of interlocutors that are not representative of the population as a whole. So these would be more marginal examples.

Now what about the second criterion "presence or absence of character strings"? That one points out something that is kind of obvious, but then again, maybe its implications are not always fully thought-through. The fact here is this: So corpora don't usually provide directly what most linguists, especially maybe cognitive linguists, are interested in. If you look at a corpus, then per se you don't necessarily find meaning in there, which obviously is something cognitive linguists would be interested in. You don't find communicative function or intention in there. You don't really obviously find information structure in there, or any obvious traces of cognition and processing, language proficiency, dialect or something—all you find is essentially character strings that either occur in certain parts or in certain locations within the corpus, or they occur in the presence or absence of other character strings.

One thing important to realize here is the very broad definition of character strings that I'm assuming here: Those can be anything. Obviously, they can be morphemes or words or constructions that, for instance, were recorded or that

### That means ...

- ... if you want to study meaning,
  - you need to qualitatively annotate (matches from) corpus data for meaning & then **correlate** the
    - the pres./abs. of certain kinds of meaning annotation w/
    - the pres./abs. of certain character strings
    - eg, annotating verbs for whether they denote literal transfer or metaphorical transfer & correlate that w/ whether the character strings around it indicate a ditransitive or a prepositional dative
      - in the part of the corpus w/ a child's utterances
      - in the part of the corpus w/ caretakers' utterances
  - eg, annotating a verb for its senses & correlate that w/ the complementation pattern denoted by the surrounding strings
    - transitive, intransitive, ditransitive, cplx transitive, ...
- ... if you want to study information structure,
  - you need to qualitatively annotate (matches from) corpus data for, say, givenness, & then **correlate**
    - correlate that w/ whether the character strings around it indicate a ditransitive or a prepositional dative

FIGURE 12

were typed at some point and that made it into the corpus, but it also includes any markup or annotation that you might have added to the corpus. So, if you part-of-speech-tagged a corpus, so that the corpus has information about, this is a verb, this is an adverb, this is a noun, and everything in there, then that of course will be represented in terms of character strings and it contributes to co-occurrence phenomena within the corpus. For instance, a certain part-of-speech tag or a certain grammatical construction tag will or will not co-occur with, for instance, a certain constructional-meaning kind of annotation or something like that. We're going to talk about much of this in more detail later.

But what that means is if you want to study meaning, you need to do some qualitative annotation of either the corpus as a whole or of matches for whatever you were looking for in a corpus for meaning. For instance, you might look for all instances of a certain verb, then you might annotate each of these occurrences of the verb for a certain sense, a verb sense that every one of these instances maybe or maybe not instantiates, and then you correlate the presence of that annotation with something else, namely, for instance, with other kinds of character strings, namely certain types of subjects that verb takes depending on what the meaning is, or certain types of grammatical tags, depending on what a meaning correlates with as a sub-categorization frame or something like that. Here's another example. You might look at verbs, and you annotate them for whether they denote literal transfer, like *He gave him the book* or

metaphorical transfer, *He brought back peace to the region* or something like that, and you correlate that with whether the character strings around the verb, in these cases, *gave* and *brought* around it sort of indicate ditransitive or prepositional dative. You need character strings to decide which construction it is—ditransitive or prepositional dative—and you have character strings on the corpus that, for every verb used, denote whether it's a literal transfer case or metaphorical transfer case, and then you can correlate them: does knowing which construction it is—ditransitive or prepositional dative—make it easier for you to predict whether it's literal or metaphorical transfer, or the other way around? That's the definition of correlation that I will assume here and in some of the other talks that follow: Correlation will be defined as knowing what one variable does, let's say constructional choice, ditransitive versus prepositional dative, does knowing what this variable does make predicting what another variable does easier? For instance, we know that ditransitives are preferred if the action denoted by the verb does denote literal transfer: Usually you would say *He gave him the book* as opposed to *He gave the book to him*, barring other information structural reasons.

Any kind of analysis that we want to do as cognitive or usage-based linguistics, if they involve meaning, they will still involve correlating presences or absences of character strings with others like that, be it words, be it annotation. The same example here: What you might do is you might annotate a verb for its multiple senses and then correlate that with the complementation patterns that you find that verb in. Certain verbs are highly polysemous, they can take on a variety of different sub-categorization frames, so you might end up with a table that says 'here are all the senses of this verb in the rows, here are all the constructions or sub-categorization frames the verb ends up with in the columns', and then you have co-occurrence frequencies that basically tell you 'this meaning of the verb obviously strongly connected to this sub-categorization frame' or something like that.

Same thing with information structure: If you are interested in the effects of givenness or newness or aboutness, something like that, and a certain constructional choice, you again would have to annotate matches or a whole corpus for givenness that you have to operationalize in some way, which is not necessarily straightforward, and then correlate them again with character strings that, for instance, indicate a constructional choice. We know that the ditransitive is preferred when the recipient is highly given: *He gave him a book*, *him* is highly given. That's what we would see in the corpus then.

Now the notion of *correlate* here is related to two implications that I think are important. One of those is something that it follows from, and the other one is one that *correlate* reflects, something that follows from that. One is this,

Note: *correlate* indicates 2 implications

- One that it follows from, that it reflects
  - the maybe single most important foundation of nearly all corpus-linguistic work, the *distributional hypothesis*
  - "[y]ou shall know a word by the company it keeps" Firth (1957:11)
  - [i]f we consider words or morphemes A and B to be more different in *meaning* than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of *meaning* correlates with difference of distribution. (Harris 1970:785f.)
  - note: these quotes talk about
    - words & morphemes
    - *meaning*
  - I prefer to interpret the distributional hypothesis more broadly: difference in function, where function covers
    - meaning, communicative function/intention, information structure, cognition/processing, etc etc

FIGURE 13

namely, what I would want to call here the single most important foundation of pretty much all corpus linguistic work, the distributional hypothesis, which is something that will be interested to discuss in another paper later. The distributional hypothesis, among corpus linguists this version of it is, I guess, the most preferred one, "You shall know a word by the company it keeps", an early citation attributed to Firth (1975:11). I have always actually preferred *this* one to it, simply because it's much more explicit and much more operationalizable: "If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution" (Harris 1970:785f). So there's a very clear statement here that, basically, presences and absences of character strings and their similarities across two linguistic elements, versus some other linguistic elements, will tell you something about how similar or different in meanings these two pairs are.

The quotes here, both of them, essentially talk about words and morphemes, so Firth (1975:11) [[said]] "you shall know a word by the company it keeps", or the Harris' (1970:785f) quote, "if we consider words and morphemes to be more similar or different than meaning" and the second thing they talked about is "meaning". Firth (1975:11) explicitly talked about meaning in the surrounding



### Note: *correlate* indicates 2 implications

- The other that follows from it, that it requires
  - *correlate* → corpus linguistics is based on correlations
    - ie the use of frequencies ( $y=0$ ,  $y>0$ ,  $y>>0$ ,  $y>x$ ,  $y<x$ ,  $y=x$ ,  $y\propto x$ , ...)
    - ie the computation of the similarities of distributions,
    - very sorry, but that means statistics
- this is a good-news-bad-news kind of situation
  - good news:
    - everything in corpus linguistics begins w/ frequencies, but that doesn't mean those are it
    - frequencies of (co-)occurrence can be used/combined in many ways to yield things of interest to cognitive linguists
    - what are those?



FIGURE 14

context here, and Harris' (1970:785f) quote even mentions it twice, "difference in meaning" and "difference of meaning". I do agree with that. But at the same time, I would like to interpret this more broadly, namely as 'difference in function'. *Function* would be basically any of the things that I mentioned before that cognitive or usage-based or linguistics in general probably that would be interested in, namely, meaning, of course I would want to keep meaning in there, but then *function* I would like to treat broader, namely also 'communicative or pragmatic function' or 'intention', 'information structure', anything having to do with cognition or processing and so on. So "distributional hypothesis", yes, but I would prefer for it to not be interpreted completely literally here only with regard to meaning, but with regard to this broader term, of *function*.

The other implication of *correlate* is a little bit more tricky for at least some part of cognitive and usage-based linguistics. This is where my tone might sometimes become slightly polemic, namely the seemingly trivial statement that *correlate* means that corpus linguistics is based on correlations. You might think "ok, duh", but there is more to it. Because first what it means is that for anything you want to look at, you need to look at frequencies of something. That's the topic that I will discuss this afternoon, a whole talk on its own. But it is worth pointing out that that means anything you want to do corpus linguistically at some level of analysis will end up with one of these statements: You have something that you're interested in, the frequency of construction,

the frequency of a sense given a certain construction, anything like that, but you will always end up with either something doesn't happen, so  $y=0$ , the frequency of this sense in this construction actually doesn't happen in your corpus sample;  $y>0$  means something does happen, there's at least one case in point of what you're interested in that was attested in your corpus;  $y$  might be much greater than 0,  $y>>0$ , that would be a mathematical way of saying 'something is pretty frequent and happens a lot'. Then of course you will have cases where you need to relate the phenomenon you're interested in to something else that it's competing with. Something like this ( $y>x$ ) might mean that one construction is less frequent than another construction, given a certain sense: 'Given literal transfer, the ditransitive is more frequent than the prepositional dative, which in turn then is less frequent ( $y<x$ )'. Something might be just as frequent as something else ( $y=x$ ), or a lot of times you will have something like this ( $y\propto x$ ), this is the mathematical notation to say 'something is proportional to something else'. As something becomes more frequent, something else becomes more frequent too, or the other way around. But anything you get out of corpora at some point will lead to something like this, even if only you look for something and you don't find it, then that  $y=0$  means there's a frequency of zero.

Secondly, what *correlate* means of course is that at some point you will compute the similarities of distribution. *Distribution* sounds like a super high-flying term here, but again, it might just be something like this: You have a comparison, a frequency of one thing, and a frequency of another thing, and you compare how similar are these numbers, are they relatively comparable? are they wildly different? and, obviously, what does that mean?

This is where for many people the horror starts, as subtly indicated by the skull here. Yes, that does mean statistics. You can't do corpus linguistics if you're not willing to engage, at least at some basic level, in statistical thinking and in statistical analysis, no matter how displeasing some people might find that. In a way, this is a good-news-bad-news kind of situation. For many people, it's just bad news, but there is good news. The good news is that, everything in corpus linguistics begins with frequencies: 'Something doesn't happen', so a frequency of zero, 'something happens sometimes'; 'something happens a lot of times', ... everything you do with corpora on some level of analysis will involve a frequency computation like that, but the interesting thing is that that doesn't mean, as it should say there, those are it. As you will hear me criticize this afternoon, there's a lot of cognitive and usage-based linguistic work that has placed a huge emphasis on the relevance of frequency as an explain-it-all kind of mechanism, but there's actually much more that corpus data have to offer, which is routinely underutilized in much cognitive and usage-based

## How are corpus data studied/explored?

- virtually all corpus-linguistic methods boil down to the following methods (or derivatives thereof)
  - **frequency list**: typically all types with their token frequencies (in descending order of frequency)
  - **dispersion**: typically a plot or a number indicating the clumpiness/burstiness of a token's distribution
  - **coll((oc|ig)a|ostruc)tion**: types and their co-occurrence frequencies (or other statistics) with some node token
  - **keyness**: types and their preference for 1+ corpora
  - **concordance displays**: the (tabular) display of instances of a node item with its contexts

FIGURE 15

linguistic work. We always start with something simple like that, but then we can build it up to talk about a whole range of other things that I will discuss sort of talk-by-talk throughout the remainder of this week. If you do it smartly, then frequencies of co-occurrence can, in fact, denote a lot of things or can be used to measure or operationalize a lot of things that are of interest to cognitive linguists. What would those be? That's something we will talk about in much more detail a little bit later.

The methods that we're using and that we're building corpora up from are the ones that are listed on this slide then. Obviously, the first and most basic thing is a frequency list. That's something we'll talk about a lot this afternoon, and then look at how to create it tomorrow morning. Essentially, for instance, all the word types might be listed in one column, like here [pointing to the table on the right side of Figure 16] and then their frequencies might be listed in a second column on the other side. This would be the top of a frequency list of the British National Corpus, a by now freely available corpus of British English of the 1990s, 100 million words. You can see that, as usual, *the* accounts for approximately 6% of all the word tokens in that corpus. That sometimes might be coupled with additional information such as part-of-speech tags. You can see that now *the*, 6 million times, that's an article; and then the preposition *of*, this many times [pointing to 2858430]; “cjc”, coordinating conjunction *and* occurs that many times [pointing at 2595716] and so on. As you all know, a lot of corpus-based work and a lot of usage-based theorizing and cognitive linguistics have placed great emphasis on the frequencies with which things

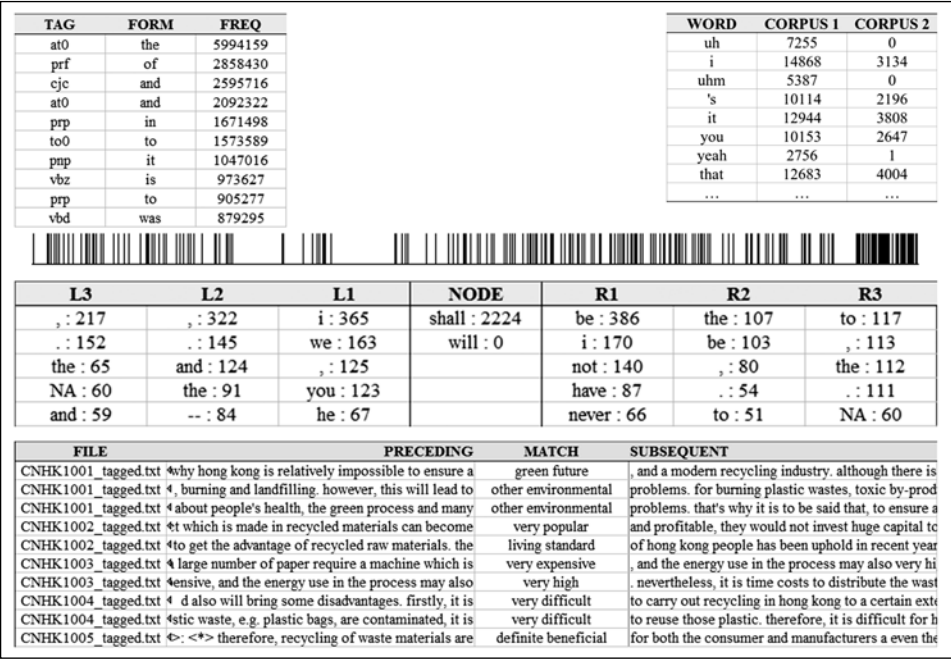


FIGURE 16

occur in corpora and what those things supposedly mean for grammaticaliza-  
tion, language acquisition, processing, and so on. That would be the most basic  
one.

The next one is, as you will hear me argue later this week, actually more im-  
portant, even though you will come up with empty hands, probably, if you look  
at most cognitive linguistic publications to even see it mentioned. That's this  
notion of *dispersion*. That's typically a plot like the one that I show here [point-  
ing to the middle bar-code graph in Figure 16], or a number that indicates the  
clumpiness or the burstiness of a token's distribution. What you should think  
of this graph as representing is this, the horizontal extent of this long line here  
is the whole corpus from the beginning to the end, so that could be as some-  
thing as small as 10,000 words, or obviously could be 100 million words. Every  
vertical line is one occurrence of the word or construction or whatever mor-  
pHEME that you're interested in. This one is interesting because, for instance,  
here this shows that there are a lot of instances of whatever this is representing  
here at the end but then there are also huge gaps here in the middle where the  
word in question is actually not attested at all. As you will see later, one of the  
points I'll be making is that this is actually more relevant than frequencies and  
I will cite some empirical evidence to bolster that claim.

There's a bunch of co-occurrence kinds of data, collocation, colligation, collocation, whatever you might want to call them. The idea being that you have, for instance, a node word, a word that you're interested in, like *shall*, and then frequencies of words around that word in different positions. The most frequent word after *shall* in this corpus is *be*; the second most frequent word after *shall* in this corpus is *I* and so on.

In corpus linguistics, we have this notion of *keyness*: How characteristic or key or important is a certain word or construction or something like that for one corpus versus another one, or for one register versus another one.

Then finally we have concordance displays. This is the display where I'm looking at these kinds of combinations here that were looked for something. Then we have the PRECEDING context and the SUBSEQUENT context. We can basically see everything that's going on in there.

Like I said before, everything you want to do with corpus data typically starts with one of these things, and typically then leads, in a second analytical step, to some frequencies based on that.

Now, the bad news is, and this is where the poison again comes in, that means there's a bigger need for proper statistical analysis than is typically recognized or even appreciated. There's a variety of relevant issues here, basically three that we kind of need to discuss. One is the role that statistical data play in a corpus linguistic/usage-based linguistic kind of setting; second, the nature of corpus data more generally and what that means for their analysis; and third,

### Note: *correlate* indicates 2 implications

- This is a good-news-bad-news kind of situation [...]
  - bad news: this means there is a big(ger) need for proper statistical analysis than is recognized, admitted, or even appreciated
- relevant issues in this connection
  - the role that statistical data play in the analysis
  - the nature of corpus data more generally and what that means for their analysis
  - the way the field is reacting to this situation
    - in general
    - with regard to particular aspects

FIGURE 17

in some sense at least to people like me most depressingly, how the field is reacting to this set of challenges or set of issues.

Here's a quote by Joan Bybee (2006:712) that I think just shows how important corpus data have become over the last 10 to 15 years. I think this is from the extremely widely-cited grammar-as-usage kind of debate with Frederick Newmeyer. It's a theoretical statement that I completely agree with [pointing to Bybee (2006:712) "it is common now to address theoretical issues through the examination of bodies of naturally occurring language use."].

But like I said, the question is how do we use these kinds of data in cognitive linguistic approaches, and to some extent, at least in psycholinguistics approaches. There's two ways you can use corpus data in work on language production, or work on cognitive linguistics, usage-based linguistics, and so on. These are those two.

You might use corpus data to constitute or define or operationalize predictors. That means you have a study that is actually not corpus-based, for instance, you're doing an experiment, but you're using corpus data in the design of the experiment. For instance, you say 'frequency is an important predictor in my study for what I want to look at; while my study is experimental, I look every word in my experiment and look up its frequency in the corpus'. Then the study is experimental, but corpus data feed into its experimental design. Here

### Corpus data in cognitive & psycholinguistics

- One can distinguish two ways in which corpus data can inform research on language production
  - to constitute/define/operationalize predictors, or independent variables (esp. in studies that might not be corpus-based otherwise)
    - e.g., when frequency data from corpora are used as statistical predictors or controls in lexical decision task studies or in self-paced reading time studies
    - e.g., when frequency and/or association data from corpora are used to construct stimuli (e.g., in multi-word unit reading time studies)
  - to constitute predictors and responses in corpus-based studies
    - e.g., in studies on phonetic reduction (based on phonetically/phonologically-annotated corpora)
    - e.g., in studies of syntactic production (based on morphosyntactically-annotated corpora)
- the former is less controversial, the latter more so

FIGURE 18

is one example: When you use frequency data from corpora to control for frequency effects in a lexical decision task, or in a self-paced reading time study. If you do a lexical decision task study and you do not control for frequency at least in some way, we'll get back to that later, you're in trouble. Second, if you look, for instance, at frequency or association data from corpora to construct stimuli, namely, when you want to measure how long does it take people to read certain multi-word units, because the idea might be, for multi-word unit which is very frequent, people read it faster, and so you want to be able to control for that.

The second kind is this one. The corpus data actually constitute predictors *and* responses in a corpus-based study. You're not doing something experimental, but you're doing actually corpus-based work in everything you look at, predictors, responses, independent or dependent variables, they are all based on corpus data. For instance, if you look at phonetic reduction in a phonologically-annotated corpus, when do people pronounce the "ing" ending as /ɪŋ/ and when as /ɪn/ or something like that or if you look at the degree of reduction of the vowel in "don't", to use a Bybee example, then that could be a study where the corpus provides both the explaining variables and the variables to be explained, the same with syntactic production: There's been a lot of alternation kinds of studies of the type that I've mentioned before, like ditransitive versus prepositional dative and so on. Obviously, there basically everything that feeds into the statistical or linguistic analysis is based on the corpus data. Now, if you compare those two things, then actually we will very quickly find that at least until relatively recently, maybe, the former kind of approach is completely uncontroversial, whereas the latter ... not so much. Most psycholinguists probably have had no problems at all to say 'we picked our stimuli or we created our stimuli in a way that was informed by corpus data', but they would have huge problems if a whole psycholinguistic or cognitive-linguistic study was based only on corpus data, especially if it's supposed to deal with matters of online processing.

Why might one be against it? Corpus data have some nasty characteristics. They are potentially too noisy, they are definitely heterogeneous, and they are unbalanced and skewed. In particular, they will exhibit a Zipfian distribution when it comes to anything happening in a slot of a construction, when it comes to words or something like that. If I remember correctly in the British National Corpus, the 100 most frequent word types constitute about 40% of all tokens, whereas the least frequent 40% of all 100 million words are just items that are used a single time. Obviously, this is as far away from nicely balanced experimental design as you can get. Also in a lot of corpus studies, predictors are highly collinear. That's something we'll talk a little bit about this afternoon. For

## Against using corpus data for both predictors & responses

- why is that? Corpus data argued to be
  - (too) **noisy**
  - (too) **heterogeneous**
  - (too) **unbalanced/skewed** (in particular Zipfian)
- in addition, corpus data often pose additional challenges
  - predictors may be (highly) **collinear**
  - **non-independence** of data points / **autocorrelation**
    - e.g., due to speakers providing more than one data point
    - e.g., due to words 'providing' more than one data point
    - e.g., due to priming effects (from many interrelated levels)
- corpus data should
  - be used for exploratory studies only
  - not be used for hypothesis-testing studies
  - "Corpora have proved useful as a means of hypothesis generation, but unequivocal demonstrations of syntactic priming effects can only come from controlled experiments" (Branigan et al. 1995:492)

FIGURE 19

instance, if you're interested in a constructional alternation, then you will find that givenness and definiteness and maybe specificity and maybe length, they will all be related and so how do you attribute causal effects to one of these as opposed to another is something I'll talk about a little bit later today as well.

There's also independence of data points or auto-correlation. Speakers obviously provide more than one data point most of the time, but also they provide different numbers of data points. One speaker will use something twice and another speaker will use the same thing but twenty times so we have a huge degree of imbalance in sort of what our input data table looks like. People have concluded from that that corpus data should really only be used exploratorily, they should not be used for hypothesis testing, and here's a quote from Holly Branigan, Martin Pickering and some other people from the 90s, (Branigan et al. 1995:492), "Corpora have proved useful as a means of hypothesis generation, but unequivocal demonstrations of syntactic priming can only come from controlled experiments." I know that Holly doesn't believe that anymore, but it was too nice a quote to not use it here.

What speaks in favor of using corpus data for this kind of stuff? Many of these things can actually be addressed pretty well. For instance, collinearity unbalancedness, autocorrelation, all those can be addressed in some statistical fashion. But there's also a much bigger advantage for corpus data that has



## In favor of using corpus data for both predictors & responses

- Yes, much of the above is true, but it is possible to address many of these challenges, plus corpus data also offer some advantages
  - **statistical control**: many of the challenges can be addressed statistically
    - collinearity unbalancedness, & autocorrelation can be addressed (remember that for later!)
  - **ecological validity** (see Jaeger's 2010 discussion)
    - experiments often introduce stimuli in balanced designs, which are not compatible with the relevant expressions' frequencies in the 'real world'
    - subjects learn distributional facts and, in particular, can do so even after only short periods of input
      - see phonological learning in Saffran & colleagues' work
      - subjects became more accepting of unconventional uses of constructions during just 8 trials (Doğruöz/STG 2014)

FIGURE 20

usually not appreciated that much, namely that of ecological validity. What I mean by that is that, if you design an experiment, then typically what you do is that we know every subject sees the same number of stimuli, every subject sees the same experimental condition, the same number of times. That's of course great, because it makes statistical analysis easy—at the same time, it's not so great because it introduces the subjects to a distribution over the time course of the experiment that is completely unrealistic because we know most things in language and slots as Zipfian distributed like that. If over the course of 30 minutes of an experiment, you expose people to a completely uniform input distribution that will have an effect. That is because subjects learn these distributional facts even over the short time course of a 20-, 30-minute-experiment already very quickly.

Here we have one example: We (Doğruöz & Gries 2014) found that over the course of just 8 experimental stimuli, subjects became more accepting of things that they usually would regard as unacceptable. At the beginning of an experiment, they were very heavily biased against a certain kind of phenomenon. Only 6 to 7 instances later, they were like, 'okay, why not? You can say that kind of'. That's just because they were exposed to these things that they normally would never hear in such short succession. You do need to control for these kinds of things, and a lot of experimental work actually doesn't go there at all.

OMG, now it's all about statistics ... :@

- Now, for many this might all seem like bad news: it seems to give an outsize role to statistics
- but it's not
  - linguistics is hardly alone in its need for proper statistical analysis, in fact
  - most cognitive/social sciences have strong statistical components
  - in fact, even many linguistic subdisciplines are highly statistical
    - psycholinguistics
    - sociolinguistics
    - phonetics ...
  - why would cognitive/usage-based linguistics be any different?
- so let us plz not respond to that with a knee-jerk reaction against empirical & statistical methods

FIGURE 21

So again, we're back to, OMG, now it's all about statistics. For many people, that's bad news, because it gives this seemingly outsized role to statistics, but it's actually not. Linguistics is not alone in its need for proper statistical analysis. Most cognitive or social sciences have strong statistical components. In fact, many linguistic subdisciplines are highly statistical, such as psycholinguistics, some areas of sociolinguistics at least, phonetics. So why would cognitive or usage-based linguistics be any different? There's really no good reason to think that. So let's not react to this with a sort of knee-jerk reaction against anything that's 'too empirical' or 'too statistical', as we will see in a moment.

Specifically, I want to discuss very briefly how really one should *not* react to this. The way I want to discuss this is by briefly talking about a recent editorial in *Cognitive Linguistics* 2016, discussing among other things a "methodological challenge". The leading article for this special issue cited a few 'attitudes', let's call it, towards the methodological challenge they see cognitive linguistics as facing. Here's a quote from the authors: "introspection and experimentation have been supplemented with corpus-based methods [yey!] and the requirement of using ever more advanced quantitative techniques risks fragmenting the field." Another quote from this article, (the three authors here, that's not necessarily *their* opinion—they're citing other people: Just to make it very clear: I'm *not* ascribing these opinions to Dagmar Divjak and Natalia and Jane), "Much of the quantitative work published under the cognitive linguistic

### How *not* to react to this ...

- A recent editorial of *Cognitive Linguistics* (2016) discussed a "methodological challenge"
  - "introspection and experimentation have been supplemented with corpus-based methods and the requirement of using ever more advanced quantitative techniques risks fragmenting the field."
  - "much of the quantitative work published under the Cognitive Linguistic umbrella does not pay enough attention to language and theory"
- rarely have I heard something more useless than
  - "concerns have been raised that the field may be becoming too empirical"
  - "numbers just for numbers' sake" (Langacker 2016)
  - "number-crunching" (Nesset 2016, Langacker 2016)
  - "empirical imperialism" (Geeraerts 2006, Schmid 2010)
- funnily enough, (cognitive) linguistics would then be a discipline where there's no 'too theoretical' but there is a 'too empirical' ...

FIGURE 22

umbrella doesn't pay enough attention to language and theory", which I've already found kind of interesting ... so "work published under the cognitive linguistic umbrella"? *Linguistic* umbrella? Cognitive linguistics does not pay enough attention to language? I find that hard to believe, because the quantitative analysis will have been done on something, and *theory*, that's an interesting point in and of itself. Of course, I can only speak for linguistics, but I wonder whether other fields have that same problem. I want to state quite clearly here that I have rarely heard something as useless as the following quotes. Something like "concerns have been raised that the field maybe becoming too empirical", I don't even know how to begin to make sense of that. Then, something like "numbers just for numbers' sake" (Langacker 2016). So with all due respect to Ron, who I do respect a lot, but this just doesn't make a lot of sense, if one looks at it in more detail. Other quotes like "number crunching" (Nesset 2016; Langacker 2016), which isn't even necessarily negative, or quotes like "empirical imperialism" by the Dirk Geeraerts (2006) and Schmid (2010). I think none of these things are particularly useful and particularly instructive. It seems like cognitive linguistics would be a discipline where obviously nothing can be 'too theoretical'—that's great—but things can be 'too empirical'—that's not great. A kind of scenario that I find extremely weird: how can linguistics as an empirical social science be become too empirical?

### How *not* to react to this ...

- And these comments are made in spite of the facts described by Dąbrowska (2016)
  - how the use of the analyst's own introspective judgments is problematic
    - "many aspects of our mental life are not accessible to introspection, [...] have to be studied using different methods"
    - "introspective judgements are influenced by our knowledge and beliefs, and often demonstrably false"
  - how even the most influential cognitive-linguistic works contain pretty much no reference to actual empirical cognitive science/psychology research (*Foundations ...*)
  - how cognitive linguistics is bad at hypothesis-testing and considering individual variation
- and at least some of these comments are by ppl who have demonstrated they lack statistical knowledge

FIGURE 23

The funny thing is that these comments are made by people in spite of a lot of facts that have been described very nicely in Dąbrowska's (2016) overview paper in that same special issue, namely how she shows that the use of introspective judgments is highly problematic. As she says correctly, I think, "many aspects of our mental life are not accessible to introspection, [...] and have to be studied using different methods" because "introspective judgments are influenced by our knowledge and beliefs, and often are demonstrably false." She says, even the most influential cognitive-linguistic works such as *Foundations of Cognitive Grammar* contain pretty much no reference to actual empirical cognitive science or psychology research, which I think obviously is highly problematic as well. Also Dąbrowska (2016) summarizes how cognitive linguistics is pretty bad at hypothesis testing and considering individual variation. She gives this example of how you talk to someone after a paper that makes a lot of reference to cognitive or processing mechanisms and you ask them about it, and then the answer would be "well, it is just a hypothesis". How about you test it? Don't just put up a hypothesis, and then at the next conference, you put up another hypothesis, but never having done the empirical work that shows whether the hypothesis actually was correct or not. The thing is that some of these comments even are by people who have demonstrated that they don't have a lot of *statistical* knowledge. Don't get me wrong; that's fine, I probably lack a lot of

## In other words, ...

- ... we need more connection to cognitive science – not less – and that entails the more rigorous empirical studies common in those areas
- there is a nice statement by Blumenthal-Dramé (2016) regarding the potential disappointment of linguists at the potential ambiguity of neurolinguistic data:
  - "this [diversity/ambiguity of neurolinguistic results] only comes as a disappointment if you presuppose that brain data should be simpler than other kinds of language-related data"
- in the present context, we can tweak this to become a good response to people scared of empirical data and their complex statistical analyses:
  - this [complexity of statistical analyses of cognitive-linguistic data] only comes as a disappointment [or threat] if you presuppose that cognitive-linguistic data should be simpler than other kinds of cognitive data

FIGURE 24

*theoretical* knowledge—we all have different strengths. But if you don't know a lot about statistics, maybe you should hold your tongue a little bit when it comes to evaluating the merits of statistical methods in cognitive linguistics.

What we need is actually more connection to cognitive science or cognitive-science kinds of approaches, which entails the kind of rigorous empirical studies that we do find in these areas. In that same really great special issue, there's a statement by Blumenthal-Dramé (2016), regarding the potential disappointment of linguists when it comes to neurolinguistic data. She writes, this (in her paper, *this* refers to the diversity or the ambiguity of neurolinguistic results). The fact that neurolinguistic results are not the big solution to all our cognitive questions, but sometimes seem to pose as many questions as we were throwing at them) diversity "only comes as a disappointment if you presuppose that brain data should be simpler than any other kinds of language-related data", right? I mean it makes total sense. But it also makes sense if you change it a bit. We make it a good response to people who are scared of empirical data and their complex statistical analyses: "This complexity of statistical analyses only comes as a disappointment or threat to you, if you presuppose that cognitive-linguistic data should be simpler than other kinds of cognitive data". If you assume cognitive-linguistic data are just as complex as many other

OMG, now it's all about statistics ... :@

- Now, for many this might seem like bad news, given the 'outsize role' given to statistics
- but it's not
  - human learning, processing, & categorization has been strongly connected to statistics
  - "In their classic review of human learning, Peterson & Beach (1967) identified that human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like  $X^2$  and  $\Delta P$  [...] & that probability theory & statistics provided a firm basis for psychological models that integrate and account for human performance in a wide range of inferential tasks." (Ellis 2006:7)
  - frequency is not just a foundational notion in statistics, but also in cognitive linguistics; the connection between the two is much stronger than superficial appeals to frequency & entrenchment suggest

FIGURE 25

kinds of data, then how would you not want to run the best statistical analysis on them?

Second, a lot of other things in cognitive linguistics or in related disciplines is very strongly connected to statistics. Human learning, processing, and categorization, all these kinds of things have been strongly connected to statistical methods. Here's a very nice quote from an overview article by Nick Ellis (2006:7), someone who I think can be universally recognized as being strong in theory *and* in methods, where he discusses the previous paper from the sixties: "Peterson & Beach (1967) identified that human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like  $X^2$ , and  $\Delta P$  and that probability and statistics provided a firm basis for psychological models that integrate and account for human performance in a wide range of inferential tasks."

So frequency is not just the foundational notion in statistics, but also in cognitive linguistics; the connection between the two is way stronger than any superficial approach that correlates frequency on the one hand with cognitive entrenchment on the other suggests.

If we relate this to some recent points or recent discussion regarding mental representation, I want to discuss two points from Dąbrowska's (2016) paper

## Relating this to some recent points: mental representation, ...

- I'd like to discuss two points made by Dąbrowska (2016) in a paper, which I appreciate *very much*
- she argues we shouldn't deduce mental representations from patterns of use
- example 1: determining at what level of granularity a construction should be posited to exist
- I agree that studying *prototypical corpus data* will probably not be able to answer *that question* - however
  - a non-prototypical corpus might
    - eg corpora w/ available auditory data
    - eg corpora w/ eye movement data
- one should be careful with statements about analysis type X since - unqualified - they might only that refer to the prototypical kind of analysis of type X
- there are many studies on mental representation issues that might very well be answerable with corpus data
  - eg the many new corpus-based studies of priming
- would Dąbrowska also say that her traceback studies do not even try to address mental representation?!

FIGURE 26

that I already cited before, a paper that you've seen I appreciate very much because of the important points that it makes.

One is this: She argues pretty explicitly that we should not deduce mental representation from patterns of use, in a sense, actually relatively related, it seems, to the Holly Branigan et al.'s quotation from before. Especially when it comes to what they said like 'causal interpretations of syntactic priming results or something like that, you can't do this with corpora, you do need experimental work for that'. So Dąbrowska here basically says, if you look at patterns of use in corpus data, you should not go to issues of mental representation from there.

Then she gives examples for that. One is this: She discusses several different constructions that are related to each other, but that are differently schematic. One of them was completely lexically filled. I think it was like a *wh*-question with an inversion or something like that. She gives one of those constructions as fully lexically filled and then she gives a variety of more abstract ones where some of the slots are just represented by a *wh*-word here, and there is a verb here or something like that. Then she says that looking at these constructions and their frequencies in the corpus doesn't really tell you which level of resolution is actually represented in the speaker's mind. The idea is you do not know whether, in *who should you ask?* or something like that, whether that specific

sequence of words in that question is mentally represented, or whether it's an instance of a more general construction, like *who should you VERB?*, that the speaker then online made into *who should you ask?*. Just by knowing how frequently does *who should you ask?* and *who should you VERB?*, if you know the frequencies of these two things in the corpus, that doesn't tell you whether the speakers have a mental representation of the more specific one or the more general one, which kind of makes sense.

I actually agree that, if you look at *prototypical corpus data*, you will probably not be able to answer that question. If *who should you ask?* is attested 15 times, and *who should you VERB?* is attested 25 times, what does that mean? It's not clear that you would say 'this one exists, and this one doesn't' or the other way around. But the critical point here is, and that's why it's kind of highlighted, *prototypical corpus data* and *that question*. Because I think she's right, but she's right only for *prototypical corpus data*, because a non-prototypical corpus might actually be able to answer that question. If you have corpora with available auditory recordings where you can measure pronunciation times, where you can measure pauses between things, where you can measure transitions between things, relate them to transitional probabilities, and so on, you might actually be able to at least have a very, very well-educated guess as to what is represented or not. Same with eye movement data. If you look at a certain eye-tracking path and you see that, sort of routinely, at the end of some expressions, the speaker backtracks to read it again, and the saccade goes backwards, that tells you something that will probably that thing wasn't represented, because if it was, the speaker would just breeze right through it. So a statement that is true of a prototypical corpus can actually not be used to say 'well, corpora in general can't handle that' unless, with all due respect to Dąbrowska, you have a too narrow understanding of what a corpus is.

Then the second thing, I don't have an example for this, but *that question*. So showing that you can't answer the question which of these different resolutions of a certain expression is represented in the speaker's mind, yes or no, or which one of these is—the fact that you can answer *that question* doesn't mean that there's a ton of other questions about mental representation you cannot answer with corpora. It is basically these two things. First, you need to be careful about any statements that you make about analysis type X like corpus data, because if you don't just talk about the prototypical analysis type X, you might actually be able to do quite a lot of things. Second, there are many studies on representation that you can answer with corpus data. For instance, studies on priming. There's now a relatively solid body of studies on syntactic priming using corpora. Priming has always been an issue about mental representation: You can only prime something if it's mentally represented in some



way. If you find robust priming data in corpora, then you have made a statement about mental representation using corpus data. In fact, Dąbrowska's own corpus-based work like her traceback acquisition kind of study, I'm pretty sure that she does want to make some claims about mental representation there, namely how children build up from simpler constructions over time get to larger constructions as they add more material as they substitute material and so on. I'm pretty sure that in her own work, there's at least a trace, no pun intended, of corpus claims leading to linguistic mental representation.

Here's a second example for that same thing. She discusses genitive marking in Polish, and she finds that there's a real big contrast between speakers' use of real words and their genitive endings in corpora and the speakers' behavior with nonce words in experiments. Ok, but first recall this notion of ecological validity: I mean corpus data are just different, they measure something else. If there's a discrepancy between corpus data on the one hand and experimental data on the other hand, it's not obvious: I mean it would need to be proven on a case-by-case basis that the experimental data are right and the corpus data are off as opposed to the other way around. Secondly, and this is, I think, even more fatal: She herself then hypothesizes where the discrepancy might come from. She says it might be due to the fact that there is a consistent usage by the

### Relating this to some recent points: mental representation, ...

- I'd like to discuss two points made by Dąbrowska (2016) in a paper, which I appreciate *very much*
  - she argues we shouldn't deduce mental representations from patterns of use
  - example 2: genitive marking in Polish, where she discusses a contrast between
    - speakers' use of real words in corpora
    - speakers' behavior w/ nonce words in experiments
  - however, first recall the notion of ecological validity
  - however, D herself then hypothesizes that the discrepancy is due to the fact that "consistent usage by the small number of speakers who have extracted the pattern"
  - in other words, *of course* corpus-linguistic methods can address this question properly, if one does it right &
    - does not just use overall corpus frequency as many do
    - uses dispersion as I have been arguing to death since 2008 ...
  - it's easy to say that method X (corpus data) can't inform Y (representation) if you don't go all the way ...

FIGURE 27

small number of speakers who have extracted a certain pattern of the genitive markings.

In other words, that actually was her admitting that corpus-linguistic methods *can* talk about this properly, namely, if you do it right: If you don't just use overall corpus frequency of something happening, but if you use dispersion as I've been arguing to death since 2008, if you figure in the fact that there's a small bunch of speakers who behave very unevenly from the rest, that means you have to make your comparison between the experimental data to the corpus data by comparing the experimental data (i) to most of the speakers that behave in some way and then (ii) the small bunch of speakers who behave in a different way, and that will already make the discrepancy at least appear in a very different light. It's easy to say that a certain method, like corpus data, can't inform representation if you don't go all the way that method has to offer. If you just go with co-occurrence or frequency, but don't figure in dispersion, then you can't really say that corpus data can't handle it—because corpus data can, if you do dispersion.

Second and slowly coming to an end here, the other thing that she's concerned with is the distributional hypothesis. That of course is essentially an extremely important point, because like I myself admitted at the very beginning:

### Relating this to some recent points: the distributional hypothesis, ...

- Dąbrowska (2016):
- she takes issue with the distributional hypothesis
  - "However, while there is no doubt that differences in meaning correlate with differences in form, distribution and semantics are just not the same thing, and the correlation is far from perfect. A particular distributional feature may provide a clue to meaning – or it may not. The assumption that differences in co-occurrence requirements always correspond to differences in meaning is methodologically useful in that it encourages us to look for such correspondences, but it may not be valid: some formal differences may be just that – formal differences."
  - true: formal differences might just be that, but
  - the distributional hypothesis – at least as I understand it – refers to more than just 'meaning' but 'function' etc more broadly, and
  - that increases the range of potentially explanatory correlations considerably

FIGURE 28

the distributional hypothesis is the foundation of pretty much anything you do with corpora, the idea being that differences in distribution on whatever level reflect, and correlate with, distributions or differences in function. What she says specifically is this [pointing to the quotation of Dąbrowska (2016) on Figure 28]. Pretty long quote, but it's relevant to discuss it here.

"However, while there is no doubt that differences in meaning correlate with differences in form, distribution and semantics are just not the same thing, and the correlation is far from perfect. A particular distributional feature may provide a clue to meaning—or it may not. [No argument there actually.] The assumption that differences in co-occurrence requirements always correspond to differences in meaning is methodologically useful in that it encourages us to look for such correspondences, but it may not be valid: some formal differences maybe just that—formal differences." I mean yes, formal differences might just be that, that's true. But first, notice how I and I think a lot of other fellow corpus linguists would understand the distributional hypothesis, namely, not just *meaning* in the sense of 'lexical, semantic meaning' or 'constructional meaning', or something like that, but something much more broadly, namely 'function'. Second, what that means is that the range of potentially explanatory correlations is much, much higher than before. If you do not find a correlation between a certain distributional formal factor on the one hand and semantic meaning on the other, then, according to my and many others' understanding of the distributional hypothesis, that's not a problem, because the correlation might be between these formal things and something else, like register or communicative function or other kinds of processing effects.

Then, she says, "even if the distributional feature does provide clues to meaning, [let's go with that for a moment] there's no guarantee that a language learner will pick up on it." That is true, too, and in fact, I will discuss an experiment in talk 8 that showcases this very clearly. But what are we supposed to take home from that? Are we now supposed to stop looking for function-form correlates because, maybe, not all of them are being picked up by a learner? Of course not. We're not going to take the fact that not every single form-meaning or form-function correlation is maybe picked up by a learner, we're not going to use this as a kill-all argument against looking for these kinds of correlations, because a lot of times we will not know which is something that will be picked up, which is something that will not be, and of course it will even vary over time.

Then, she says "distributional features are often correlated with each other". Remember the notion of collinearity before? What was the example I gave? Givenness and definiteness, maybe specificity, and length will all be related if you look at them in corpus data. "So distributional features are often correlated

## Relating this to some recent points: noticing & using correlations, ...

- Dąbrowska (2016):
  - "even if a distributional feature does provide clues to meaning, there is no guarantee that a language learner will pick up on the feature"
  - that is true, too, but so what? are we now supposed to stop looking for form-function correlates because, maybe, not all of them are picked up by the learner?
  - distributional features are often correlated with each other, and most of the statistical models that corpus linguists use (e. g., regression) assume that the predictors vary independently. As a result, analysts have to choose one of a set of correlated features, and the decision which one to include in the model and which to leave out is fairly arbitrary."
  - this is simplistic and misleading
    - if predictors A & B are highly correlated (eg  $R^2=0.9$ ), then
      - choosing A over B or vv is arbitrary & inconsequential
      - there're better ways to deal w/ collinearity than 'A or B'

FIGURE 29

with each other and most of the statistical models that corpus linguists use (e.g., regression), assume that the predictors vary independently. [That is correct.] As a result, analysts have to choose one of a set of correlated features, and the decision which one to include in the model and which to leave out is fairly arbitrary." So the idea being you want to look at the distribution of two constructions, you have five predictors here that are supposed to explain what happens here when the speaker uses this and when the speaker uses that, and so what you're saying is that, of these five things, these three are related. Then, she says analysts have to choose one of a set of correlated predictors. So she says, basically, of those three then the analyst has to choose, let's say, the second one and correlate that with that, the choice is arbitrary.

Again, with all due respect to Ewa Dąbrowska and everything, but that's just, I call it *simplistic* and *misleading* here. Strictly speaking, it's actually wrong for two reasons. One is if these three predictors are highly correlated, like *really* highly correlated, then which one you choose is arbitrary, but it's also inconsequential. Because if these three things are super highly correlated, then picking this one for whatever reason, if they are so highly correlated, then whatever this one tells you will also cover what these other two things do. It's like debating how to measure the length of a noun phrase. You can measure it in characters or in morphemes or in words. When you pick one of those three

### To wrap up

- Corpora come in many shapes, traditional/prototypical ones, but also more marginal ones that extend the range of issues that corpora can address
- don't overgeneralize what corpora can(not) do just on the basis of the prototypical corpus (study)
- corpus data are frequencies of character strings - original source text & annotation - which need to be correlated with other things; for that we need statistics
- that is no need to run & hide from the imperialists to a qualitative or theoretical safe space - that's a reason to learn
  - how to assess when & how corpus statistics can help
  - how to do them properly to address the challenges that come with high(er) ecological validity
- only then can we assess the role and relevance of quantitative corpus data for particular questions

FIGURE 30

'arbitrarily', that number will still cover what the other two are doing to an extremely high degree.

Secondly, it's actually not even true to say that you *have* to choose one of the three. There *are* statistical methods that allow you to do much better than that. I don't want to turn this into a statistics seminar, but there's ways in which you can merge the information of these three together into one new variable that then covers all of what those do. The assumption that, or the implication here that you have to pick one is simply wrong. If you know your statistical stuff, you know that's not what you need to do.

All right, to wrap up, corpora come in many shapes. We have traditional and prototypical ones that usually constitute spoken or written data. [[Corpora]] come in some forms of textual format in Unicode and something like that. But there's also a range of much more marginal ones and that is sometimes very tricky, because the kind of requirements they pose on your data handling abilities and your statistical analyses are more challenging. But at the same time, it's great because they increase the range of issues you could address. Maybe you cannot make claims about mental representation from a traditional interview transcript, but you probably can from an eye-movement corpus. If you think about using corpora, or if you read about other people's thoughts of using corpora, again bear in mind corpora are prototype category. Just because

people say corpora cannot do X doesn't mean that *all* cannot: they might be right in that the prototypical corpus cannot but that doesn't mean that if you were to use other kinds of corpora, that would be that they can then still not study X. So don't overgeneralize with regard to what corpora can or cannot do, just on the basis of a prototypical corpus.

Then third, corpus data are frequencies of character strings. That's even true of eye-movement data because those will be represented in some sort of file type or format. These character strings are original source text—what was originally said or spoken or written—plus also the annotation, and anything you do from corpus data will consist of correlating text and annotation, either yours or the one that comes with the corpus. In some way and for that, you will need to use statistics in some way. There's just no way around it, whether you like it or not. That doesn't mean there's a need to run and hide from some imperialists to some qualitative or theoretical safe space where you can theorize without any regard to what does it actually mean, "empirical". It just means that you need to learn to assess when and how and of course also what type of corpus statistics can help—I'll talk about that a lot over the course of this week—and how to learn them, how to do them properly and address the challenges that come with the higher degree of ecological validity. Experiments are great at one thing: They're great at sort of holding everything constant, giving you a certain set of conditions and, typically at least, they're extremely easy to analyze statistically once you know a little bit. Corpus data come with all the noise and heterogeneity that they come with, but they have that higher degree of ecological validity you do not need to control as much, for instance, within corpus learning effects as you would need with an experiment. Because only then can you address and assess the role and relevance of corpora for particular questions that you might have in cognitive or usage-based linguistics. Thank you.

## On—and/or Against—Frequencies

Welcome to the second talk. As you can see from the title and as I mentioned this morning, it will be about frequencies and it will maybe be a less strong endorsement of frequencies as they are currently used a lot of times in cognitive and usage-based linguistics as one might expect in a forum like this.

As I said this morning, corpora do not really provide anything of cognitive linguistic interests per se; basically all the things that you want to measure you have to measure in terms of something with character strings, so things like meaning, communicative function, intent, information structure, all of these kinds of things, they all have to do with the presence or the absence of certain

### A very brief recap

- I mentioned earlier today that corpora do not really/directly provide what most (cognitive) linguists are interested in
  - meaning, communicative function/intention, information structure, cognition/processing, language proficiency/dialect, ...
- they provide information on the presence or absence of character strings
  - in (certain parts/locations of) corpora
  - in the presence or absence of other character strings
- note: character strings can be anything: morphemes, words, constructions, anything including and especially markup/annotation!
- in this talk, we will focus on the first type: frequency in (certain parts/locations of) corpora - later talks will focus on the second type

FIGURE 1



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611207>

character strings either in certain parts or locations of the corpus, which is something we will talk about tomorrow, or in a corpus in general, and then also in the presence or the absence of other character strings, where the important point that I wanted to drive home this morning, one of those, was that character strings can be anything, they can be the actual source text that was written or spoken or signed or whatever, but also any kind of markup, any kind of annotation that you might have added during later analytical process.

So what I want to talk about in this talk here now is this first part of the concerned with the presence or the absence of character strings. So I want to talk about frequencies of things either in corpora in general, or in certain parts, or in certain locations in corpora. Later talks will be concerned with what happens in the presence or absence of other kinds of things, so later talks will be concerned with conditional probabilities or contingency or association depending on how you might want to call it.

Obviously, there are different kinds of frequencies we need to distinguish and literature in general does distinguish and this part here is going to be very trivial in a sense probably for most of you. The first kind of frequency distinctions you might want to make is that between token frequency and type frequency, where, obviously, again just to get that all on the same page, I know this part is boring, token frequency is concerned with the numbers of times a

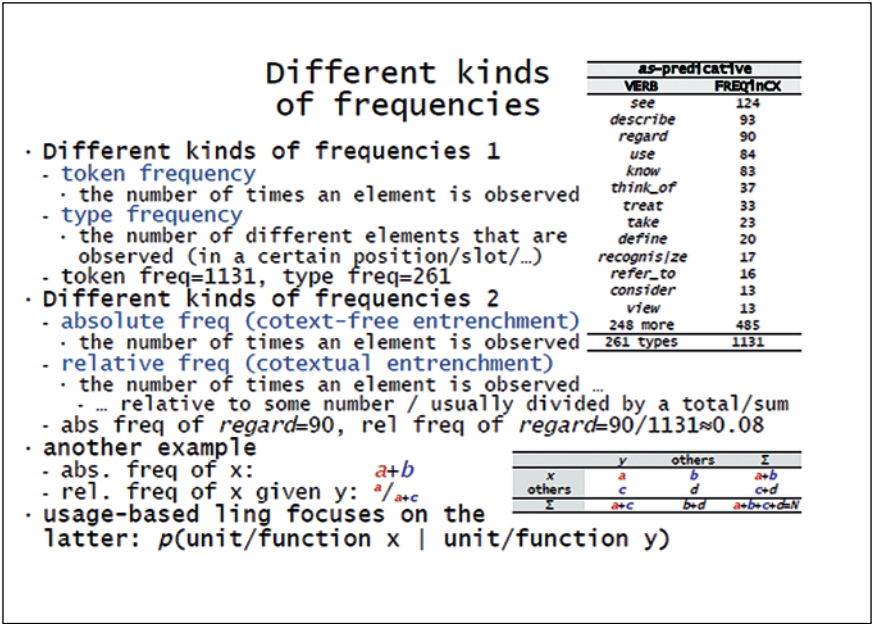


FIGURE 2



certain element is observed in the corpus and certain part of a corpus or, even perhaps, in a particular constructional slot or something like that whereas type frequency is concerned with the number of different elements that are observed either in a corpus in general, in a certain part in the corpus, or in a certain constructional slot that a corpus might have.

To give you one example here, with a construction will be talking about quite some time, or quite repeatedly, here you have a frequency list. The construction in question is the *as*-predicative as Beate Hampe, Doris Schönefeld, and I called it. That is the construction that is often observed with, as you can see, verbs like *regard*, or *see*, or *describe*, things like that. It's the construction that you have an example like *He regarded himself as a cognitive linguist*, *He saw himself as a usage-based linguist*, *He described himself as a usage-based linguist*, something like that. So basically, we have a verb, we have a direct object, we have an *as*, and then we have a complement of the *as*, and a lot of times it's a noun phrase, but it can actually also be a non-finite verb phrase, it can be a prepositional phrase, or anything like that. But again, the construction is something like, especially in the passive, *He was regarded as one of the most famous linguists in the world* or something like that, for instance.

If you look at this construction in the International Corpus of English, the British component, you would come up with this frequency list of the verbs in the verb slot. You can see that altogether we have 1,131 instances of that construction. So that's the token frequency of that construction, and it's the token frequency obviously of all the verbs that figure in that construction at one point or another. Then, the number of different verbs that are found in that verb slot is 261. So on average, every word occurs five times or something like that, but of course the distribution is very Zipfian so there is a few verbs that account for vast majority or for at least a large proportion of the items and then it levels off pretty quickly. Even after only ten verbs or so we get down from 124 to only 13, and then the remaining 248 types account only for not even twice as many verb tokens. The remaining 248 ones are attested only twice on average which of course contains a few that are attested ten times or eight times or something but then a lot of hapaxes, so words, verbs, that show up only a single time in there.

Then with those definitions here, again the token frequency of the construction is this (1,131). The type frequency of the verbs in the verb slot is that (261). Again, probably not particularly sensational news here, but just so that we get on the same page.

Then, a different kind of distinction of different frequencies would be that of absolute and relative frequencies. So an absolute frequency would be the number of times certain element is observed. In some of the cognitive linguistic

literature, particular that by Hans-Jörg Schmid, absolute frequency has been equated with, I guess, what he called context-free entrenchment. As you will see in a moment, that's the general tendency to equate frequency with entrenchment or at least operationalize it, it is not always clear. The absolute frequency would be context-free because it's not dependent on a certain context—it's just the overall frequency of something occurring in a corpus, for instance. Then, the relative frequency would be co-textual entrenchment, so that is a frequency that depends on the certain context. Namely, you look at something relative to some other number or usually divide it by some kind of total sum. In this case, we would say, for instance, the absolute frequency of *regard* in this construction is 90, which is what you see up here, and then the relative frequency of *regard* would be the absolute frequency divided by the overall total, 90 divided by 1,131, that's 0.08, or 8%, if you will. So again, relatively straightforward.

Another way in which this is often talked about, in particular in studies that later have to do with association or contingency or any kind of collocation, co-occurrence phenomena is using this kind of 2 by 2 table. So here we have two linguistic elements, whose concurrence we're interested in. We have an element called *x*, *others* means 'every element that's not *x*', so this could also say *x*: yes and *x*: no. Then we have a second element that might be called *y*, and again we have *y*: yes, that happens, or *y*: no, it doesn't. Crucially, for things that we will be discussing later, *x* and *y* can be of the same level of linguistic organization, so for instance, these could both be words, right? They could also be, for many examples to be discussed and like here, it could be that one of them is a word, yes or no, and the other one is a construction in which the word occurs, yes or no. In that case, the frequency schematically denoted here as *a*, that's the frequency that the verb *x* occurs in construction *y*. In this case, the absolute frequency of *x* would just be the row total. *X* occurs this many times with or in that construction and these many times elsewhere, so its overall frequency is the sum of those two.

The same thing for the relative frequency. The relative frequency of *x* given *y*, is, when *y* is there, there is *a+c* items and *a* out of *a+c* items are actually when the two co-occur together, so that would be that. Again, I mean this is just schematic representation of what are just fractions or percentages or whatever you want to look at. I think, in general, it's fair to say that in usage-based linguistics, the relative frequency of things is more important in general at least than the absolute frequency. So we're looking at what I've called conditional probabilities, the probability of some unit (like a word, a morpheme, or some function, certain communicative function or certain discourse structural function) given (so that vertical line here, |, that pipe, as it's called, means 'given') some other unit or some other function. This here, *a* divided by *a+c*, that would

## The role of token frequency

- For a long time, token frequency has been an important **cause** or **control** in psycholing studies
- it has a statistically reliable **effect** on naming (Oldfield & Wingfield 1965, Lachman 1973; Lachman, Shaffer, & Hennrikus 1974, and many more)
- it seems to **affect** reaction times in lexical decision tasks, word naming, picture naming, etc (Howes & Solomon 1951, Forster & Chambers 1973, ...)
- it seems to **affect** phonological more than semantic retrieval (Huttenlocher & Kubicek 1983, Kelly 1986)
- but it became relevant for
  - (theoretical) linguistics in general
  - cognitive linguistics/usage-based ling in particular
- via the development of cognitive linguistics, cognitive grammar, usage-based linguistics, & cognitive (Goldbergian) construction grammar
- via the usage/grammar debate at the LSA and in *Language* (Bybee 2005 / Newmeyer 2005)

FIGURE 3

be the presence of  $x$  given that  $y$  is there. Again,  $a$  divided by  $(a+c)$  means given that we have this many occurrences of  $y$ , how many of those are also  $x$ , so that this ( $a$ ) is divided by that ( $a+c$ ). Most of the time, that's what is more of interest in these kinds of applications.

What's been the role of token frequency? For a long time, it's had an extremely strong role in especially psycholinguistic studies, namely, as either an actual cause or something, so it was supposed to be the reason for something else happening, or at least minimally as a control for something else. So you did an experiment on reading times, for different degrees of abstraction of nouns or something, but what you wanted to control for is the effect that is not also due to frequency, so you held frequency constant, for instance, at least people tried.

Either one of those two were extremely frequent. Why is that? Because there is a ton of studies that have shown a few things, such as that it seems to have a statistically reliable effect on naming. I am highlighting *effect* in blue to indicate that that kind of language is causal language, right? If you say  $x$  has an *effect* on  $y$ , then you are not just saying they are correlated or co-vary—they say that this thing causes that other thing.

It also seems to affect reaction times in lexical decision tasks or specifically word and picture naming. Again, notion of affect reaction time seems to

## The role of token frequency with regard to entrenchment

- In CogLing, token freq became assoc'ed w/ entrenchment ·

"continuous scale of entrenchment in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence." (RWL 1987:59)

"[t]his seems highly convincing, not least in view of the considerable body of evidence from psycholinguistic experiments suggesting that frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse [...]. As speed of access in, and retrieval from, the mental lexicon is the closest behavioural correlate to routinization, this indeed supports the idea that frequency and entrenchment co-vary." (Schmid 2010:115f.)

- note:

- causal language in Langacker: *depending*
- causal & correlational language in Schmid (2010):  
*determinant* but *co-vary*

FIGURE 4

suggest that causal combination here. It seems to affect phonological more than semantic retrieval, but the thing is for us, now as cognitive or usage-based linguistics, it became relevant in particular via first of course the development of cognitive linguistics, cognitive grammar, usage-based linguistics, maybe Goldbergian construction grammar, kind of along the lines of what Arie mentioned this morning. Of course also, I mean, at least in the US in particular because of this usage-versus-grammar debate at the Linguistic Society of America meeting, and then sort of dueling articles in the journal *Language* between Joan Bybee (2005) on the one hand, sort of on the usage-based side and Frederick Newmeyer (2005) on the more generative/structural side of things.

How did that affect things in cognitive linguistics? Well, it became associated with this notion of entrenchment. Here are some quotes that I think make that point very clearly: The first one from *Foundations I*,

continuous scale of entrenchment in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence.

LANGACKER 1987:59

Then much more recently a quote from someone who has worked a lot on entrenchment, Hans-Jörg Schmid again:

[t]his seems highly convincing, not least in view of the considerable body of evidence from psycholinguistic experiments suggesting that *frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse* [...]. As speed of access in, and retrieval from, the mental lexicon is the closest behavioural correlate to routinization, this indeed supports the idea that frequency and entrenchment co-vary.

SCHMID 2010:115f.

Probably sounds familiar to many of you, but the one point that I do want to highlight here, just to be pedantic, you know: We have causal language here indicated in blue, for instance, in the Langacker quote, the word *depending* to me at least suggests causal attribution not just correlational talk and then in Schmid we have causal and correlational language. If you use the word *determinant*, I mean that's pretty much making a causal claim, right? Something determines something else. If you say two things *co-vary*, you are kind of dodging that question. I mean maybe with good reason, you know, nothing bad about

## This is what token freq/entrenchment has been argued to do or be related to

- **Learning & acquisition**
  - token frequency counts how often a particular form appears in the input. (Ellis, Römer, & O'Donnell 2016)
  - ease/earliness of acquis. (Casenhiser & Goldberg 2005)
- **predictability, ease and speed of access**
  - psycholinguistic studies and Schmid as quoted above
- **routinization, reduction, language change, development of new forms** (Schuchardt 1885, Fidelholtz 1975, Bybee & Thompson 1997, Gómez 2002, Onnis et al. 2004, Aslin & Newport 2012, ...)
- **categorization/category formation**
  - exemplars presented w/ higher frequency are classified more accurately & are seen to be more typical
  - exemplars that are more similar to the higher frequency items are classified more accurately & are seen to be more typical (Ellis, Römer, & O'Donnell 2016:60f.)

FIGURE 5

that, but still it is worth to point out that this is a statement, this is language that has to do with correlation, but *co-vary* does not necessarily indicate a causal relationship there.

Now what's token frequency and entrenchment then supposed to be doing? That's a ton of findings, many of which will again be familiar to you. Obviously, for instances, in learning and acquisition, we've seen that token frequency counts how often a particular form appears in the input, and so obviously it will have an impact on which words, which constructions, and which combinations of things will be attested, this being a quote from a recent monograph by Ellis et al. Obviously, token frequency and entrenchment have to do with ease/earliness of acquisition: Things that a child hears more often, more early in his or her life will lead to higher rates of acquisition of those linguistic science. We've seen that it has to do with ease and speed of access, it also has to do with predictability, I mean entrenchment, that is. So, we have a bunch of psycholinguistic studies that were cited on the previous slides and then the same quote from Schmid that I've just read out to you.

Obviously, frequency and entrenchment have something to do, from a usage-based linguistic perspective, with all these things, with language change and development, so routinization, reduction, language change and development, are all minimally correlated with, if not determined by, this. So for instance, reduction: high frequency of co-occurrence might lead to phonological reduction, remember that famous study Bybee showing that the [o], what is represented by the letter *o* in *I don't know* is pronounced less strongly than if the word after *don't* is less frequent. For instance, obviously, routinization will be strongly correlated with high frequency of occurrence things together, because Bybee would say at least to chunking and all these kinds of things.

Then, there is supposed to be a relationship between token frequency and categorization or category formation. That's because if you have exemplars of a category that are displayed with a higher degree of frequency, then they are classified more accurately as being a member of that category. They are all considered to be more central members of that category so there is a kind of prototype effect there. Another effect would be that exemplars that are more similar to high-frequency items are also classified more accurately and are seen to be more typical.

Something can be considered more prototypical because it's more frequent in and of itself, or because it is very similar to something that is very frequent in and of itself, but both obviously are related to frequency effects.

How was this connected then to earlier psycholinguistic models? One again relatively long quote, but I think it's instructive for two talks that follow later, so here's a quote from Nick Ellis's (2006:9–10) discussion:

We are more likely to perceive things that are more likely to occur. The power law of learning describes how the resting levels of detectors for words, letters, and other linguistic constructions are set according to their overall frequency of usage so that less sensory evidence is needed for the recognition of high frequency stimuli than for low frequency stimuli. Each time we process a stimulus, there is a practice increment whereby the resting strength of its detector is incremented slightly, resulting in priming and a slight reduction in processing time the next time this stimulus is encountered.

This is important for things that follow later because the idea is that it basically assumes a kind of logogen or whatever, interactive activation kind of model of the mind, where for instance, words, or constructions or things like that are nodes and if you process a certain word, a certain node, if you activate it more often, then its resting level of activation will slowly be increased so that it's easier to activate later, which would of course be a sort of very rational, or sort of evolutionarily useful process because it means that things that you hear often can be activated quickly without much cognitive effort, so to speak. It will be apparently later why I am riding on this here.

### How was this connected to (earlier) psycholinguistic models/work?

"We are more likely to perceive things that are more likely to occur. The power law of learning describes how the resting levels of detectors for words, letters, and other linguistic constructions are set according to their overall frequency of usage so that less sensory evidence is needed for the recognition of high frequency stimuli than for low frequency stimuli. Each time we process a stimulus, there is a practice increment whereby the resting strength of its detector is incremented slightly, resulting in priming and a slight reduction in processing time the next time this stimulus is encountered."

(Ellis 2006:9-10)

"Practice promotes proficiency (eg, Anderson, 2009; Bartlett, [1932] 1967; Ebbinghaus, 1885). Learning, memory and perception are all affected by frequency [...]: The more times we experience something, the stronger our memory for it, and the more fluently it is accessed."

(Ellis, Römer, & O'Donnell 2016:45f.)

FIGURE 6

## How was this connected to (earlier) psycholinguistic models/work?

- Comprehenders know the relative frequencies with which individual verbs appear in different tenses, in active vs. passive structures, and in intransitive vs. transitive structures, the typical kinds of subjects and objects that a verb takes, and many other such facts. This information is acquired through experience with input that exhibits these distributional properties. A verb's behavior is also closely related to its semantics and other properties specific to it ... this information is not some idiosyncratic fact in the lexicon isolated from "core" grammatical information; rather, it is relevant at all stages of lexical, syntactic and discourse comprehension. (Seidenberg & MacDonald 1999:579f.)

FIGURE 7

The same thing here from a more recent monograph summary, “practice promotes efficiency” (e.g. Anderson, 2009; Bartlett, [1932]1967; Ebbinghaus, 1885). We all know that. Then, “learning, memory and perception are all affected by frequency. The more times we experience something, the stronger our memory for it and the more frequently it is accessed” (Ellis, Romer, & O'Donnell 2016:45f), relatively uncontroversial probably in most circles.

Another quote that really highlights very strongly this psycholinguistic connection—and this is where the slides that you have in the program book will differ slightly from the ones I'm presenting here, I've added a few things yesterday—is this: (relatively old, but still a very nice psycholinguistics paper):

Comprehenders know the relative frequencies with which individual verbs appear in different tenses, in active vs. passive structures, and in intransitive vs. transitive structures, the typical kinds of subjects and objects that a verb takes, and many other such facts. This information is acquired through experience with input that exhibits these distributional properties. [I guess we would not disagree with that.] A verb's behavior is also closely related to its semantics and other properties specific to it ... this information is not some idiosyncratic fact in the lexicon isolated



from “core” grammatical information; rather, it is relevant at all stages of lexical, syntactic and discourse comprehension.

SEIDENBERG & MACDONALD 1999:579f

In much of that literature, token frequency really kind of has to shoulder a lot of things, because it is supposed to be a cause or explanandum of so many things.

Now what about type frequency, the other kind of frequency we talked [[about]] at the beginning? Type frequency has also been an important cause or control in a variety of studies but different kind of things, in particular, type frequency has played a role in everything having to do with morphological or lexical or syntactic productivity, because, for instance, probably most famous studies in this area, Bybee & Thompson (1997), Bybee & Hopper (2001), productivity of patterns is a function of type, rather than token, frequency: The more items you see in a certain position in the construction, like in the verb slot or in some other complement slot or something like that, the less likely the construction is associated with a particular item, and the more likely it is that a general category is formed over the items in that production. So the idea being, if you have a certain word and after that word there is only one other thing that can occur, then this is not a general construction, then this is

### What about type frequency?

- For a long time, type frequency has been an important cause or control in studies of
  - productivity
    - productivity of phonological, morphological, & syntactic patterns is a function of type rather than token frequency (Bybee & Thompson 1997, Bybee & Hopper 2001)
    - the more items in a certain position in a construction, the less likely the construction is associated with a particular item, & the more likely it is that a general category is formed over the items in that position
    - the more items the category must cover, the more general are its criterial features and the more likely it is to extend to new items
    - high type frequency ensures that a construction is used frequently, thus strengthening its representational schema
  - grammaticalization
  - category formation
- and it might be a decent indicator of phraseologisms
  - hermetically \_\_\_\_\_

FIGURE 8

maybe a multi-word unit and that's kind of it. Whereas if you have a word and many other things can occur within one syntactic frame, then the assumption would be that this is probably a more general kind of pattern. There are a lot of things that can happen after *the*, but they are [[often]] nouns, so maybe there is a thing like a noun phrase. There are other expressions that do not allow a lot of variability in front of them or after them—I will show you an example in a moment—and so those are probably not general constructions or general categories.

That of course also is what's then sometimes used in a discussion of things like semantic bleaching and whatever: The more items a category must cover, so the higher the type frequency of things that are going to a certain slot are, the more general are its criterial features, so the more abstract or the more schematic they are, and the more likely is that to extend to new items, so to permit other things to go into the same slot, to the degree that semantics and everything are compatible.

Of course, high type frequency ensures that a construction is used at least somewhat frequently: You can't have a high type frequency and a low token frequency, because each type has to be attested at least once and that of course means there is also a connection to strengthening or entrenching a representational schema.

Similar arguments then go for grammaticalization and category formation. I do think that, because of what I just said before, the kind of logic discussed here, that in some cases at least type frequency might be a decent indicator of phraseologisms. If you take this word, then there is a slot after it and in English you will find pretty much only a single word after *hermetically* which is *sealed* [*hermetically* \_\_\_\_]; there is nothing else. So no matter how often that word happens, the type frequency after it, on the whole, will be close to one. So that indicates that *hermetically sealed* is probably going to be mentally represented, notwithstanding any discussion of Dąbrowska and myself this morning.

Token frequency, then, has been one of the most widely-used predictors in cognitive/usage-based linguistics. In the list of applications that I showed you before, I mean that was everything, that was category information, learning, acquisition, processing, either in speed of access, everything and their mother was somehow related to frequency of (co-)occurrence in the corpus. In a way that's good, because we *have* seen a bunch of empirical studies that have shown that there is high degree of correlation at least between frequency and occurrence or frequency and co-occurrence. In fact, in a very recent and very psychologically informed overview monograph, Chater & Christiansen (2016) argue that “contemporary theories of perception and action have proposed that the cognitive system aims to build a probabilistic model, which captures

Ok, but ...

- Token frequency has been one of the most widely-used predictors in cognitive/usage-based linguistics - that's
  - good
    - many cognitive & (psycho)linguistic phenomena are correlated with frequency of (co-)occurrence
    - "contemporary theories of perception and action have proposed that the cognitive system aims to build a probabilistic model, which captures the statistical structure of the external world" (Chater & Christiansen 2016)
    - we seem to be honoring Lakoff's (1991) cognitive commitment
  - understandable: corpora provide (relatively) easy access to many kinds of frequencies of (co-)occurrence
- but there's also cause for concern
  - is especially token frequency a cause or a correlate of cause(s)?
  - much usage-based work seems to assume the former but does not always bother to carefully separate both, or explore other options ...

FIGURE 9

the statistical structure of the external world", basically saying that in some way, kind of like the Ellis' quote from this morning, in some way, we are trying to build a statistic model, keeping track all sorts of co-occurrence structure in the input and around us.

Of course, it also seems as if it makes us honor Lakoff's (1991) cognitive commitment, because we are looking sort of at neighboring discipline like psycholinguistics and say, "Look! Frequency does all these things. Isn't that great? Wouldn't it be great if what we are looking at, the effects are compatible with what we see in the empirically often much more rigorous work over there?". It's also understandable that frequency becomes this widely used predictor because you can get it relatively easily. Even if you don't like anything of corpus linguistics, there are unfortunately some corpora with too-easy access available out there, that allows you to get access to frequency data very quickly, so you can pretend for a moment that that is a real corpus-linguistic study, and you have some frequency data to work with.

There is also some cause for concern, I think. One is this, namely, the question of whether token frequency is actually a cause itself. I mean is it really the reason for something or it's just correlated with one or more causes. The thing is that much usage-based work seems to assume the former, but does not actually even bother much, it seems, to carefully separate the two options. I don't see a lot of usage-based linguistic studies that says 'ok, frequency could be the

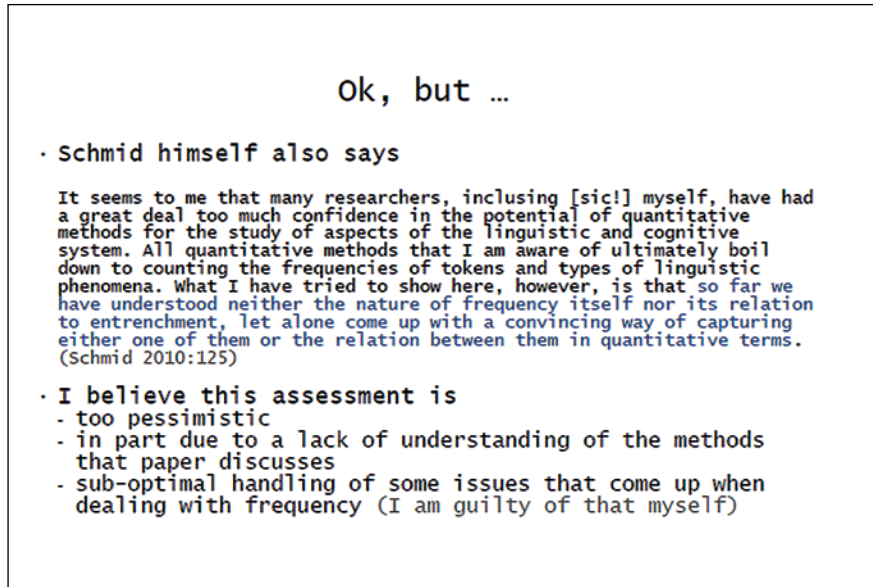


FIGURE 10

reason here, but let me test three other things that behave similar to frequency and then let's see what is really going on'. That kind of work has been done in some psycholinguistic studies, but in cognitive linguistics, I don't see that very much, also not in the theoretical work, as you saw in the quotes before, Langacker, Schmid, and others.

If you look at what Schmid himself has to say about this, he says, "many researchers, including myself, have had a great deal too much confidence in the potential of quantitative methods [for the study of aspects of the linguistic and cognitive system]. All quantitative methods that I am aware of ultimately boil down to counting the frequencies of tokens and types of linguistic phenomena. [I agree, but then and now the blue part] so far we have understood neither the nature of frequency itself nor its relation to entrenchment, let alone come up with a convincing way of capturing either one of them or the relation between them in quantitative terms." (Schmid 2010:125) I think it's a little bit too pessimistic, this kind of assessment and I am sorry to say but in part, I think it's because at least the methods he discussed in that paper fall short of showcasing the whole variety of things one should be looking at.

Third, the way he deals with frequencies in some of the empirical studies is not ideal itself, either. I think part of the pessimism basically needs to be written down to *that* kind of problem, not to an actual problem in relating frequency and other kinds of things as you will see later.

## So what is one to do?

- Obviously one can try and improve the "quantitative terms" that are used to understand the relation of frequency and entrenchment
  - yes, pretty much everything does indeed "boil down to counting the frequencies of tokens and types of linguistic phenomena", but that doesn't mean one can't add additional layers of sophistication
    - what things are counted
    - where things are counted
    - how counts are related to each other & transformed & ...
- more importantly, however, one can question the straightforward relation between frequency & Y:
  - frequency and the above-mentioned (linguistic) phenomena
  - frequency and entrenchment even more generally

FIGURE 11

How do we handle that, what we are supposed to do? So one thing you can obviously do with them is trying and improving that, these *quantitative terms*, as he calls them, to understand the relationship between frequency and entrenchment better. Pretty much everything does boil down to counting tokens and types, but that doesn't mean that it stops there. Like I said this morning, 'just because you begin with frequency, doesn't mean that's it'. There are a lot of different things can be done with frequencies if you know how, and remember the Dąbrowska scenario: Sometimes you might find the ways in which frequency data from corpora are not corroborated by experimental data, but that's only because you haven't added dispersion—distribution within the corpus—to the equation, so everything is off because of that.

What we need to look at is exactly what do we count and where do we count it, and how do we relate different kinds of counts to each other and maybe transform them in such a way that they inform our analyses of frequency better. But there is another way one can proceed as well, something else one can do on top of this, namely, you can just question the straightforward relationship between frequency as a cause of, and I am using *Y* here as sort of 'any other effect', anything that you want to explain with frequency. Basically, this is sort of maybe not the most elegant way of saying, 'maybe the cause-effect

But why would one do the latter,  
given all the empirical support?

- Because monofactorial studies have virtually nothing to contribute to usage-based (corpus) work (there, I said it!)
- for the sake of simplicity, let's explore this with a very mundane non-linguistic example, the efficiency of cars
  - measured in mpg, miles per gallon
  - which is inversely related to liters per 100 km
    - the higher the mpg, the more efficient a car
    - the lower the mpg, the less efficient a car

FIGURE 12

relationship between frequency and something else is actually not as strong as has been assumed for such a long time now'.

So that can be applied on two levels. The first can be, the one I just mentioned, namely, you look at frequency as a cause, and then maybe some linguistic phenomena: routinization, earliness of acquisition, something like that. All you can just in general say frequency and entrenchment maybe they are not causally correlated in the way that many people have been talking about so far.

That of course leads to some tricky questions. One of them especially would be why would you do that? Given that there seems to be such a strong amount of empirical support. Why would you question that strong relationship between frequency as a cause of something else? The main reason for that is this, I know that is a strong statement, but I still believe it is true, namely those monofactorial studies, so studies that postulate a relationship between one cause be a frequency of something else—here, I am talking about frequency, but it is true in general, I think—so monofactorial studies postulating a cause-effect relationship between one thing and one thing else have not much, if anything, to contribute to usage-based corpus work, I think. Like I said, I know it is a strong claim, but I want to talk a little bit about it and I am going to

## What affects the efficiency of cars (measured in mpg)?

- In other words,
  - you're pretending we know nothing about mpg already
  - you're leaving all mpg variability up for grabs by disp
- but that's delusional/too generous: you already know that cylinder, horsepower, & weight affect mpg

```
> summary(prior.know1 <- lm(mpg ~ (cyl+hp+wt)^2, data=mtcars))
[...]
```

```
Multiple R-squared:  0.895,    Adjusted R-squared:  0.8697 ***
```

- option 1: you need to test whether disp adds to what we already know

```
> summary(real.test.of.new.hyp <- lm(mpg ~ (cyl+hp+wt+disp)^2, data=mtcars))
Multiple R-squared:  0.896,    Adjusted R-squared:  0.8464 ***
```

```
> anova(prior.know1, real.test.of.new.hyp, test="F")
```

```
Res.Df  RSS Df Sum of Sq  F Pr(>F)
2      21 117.16   4      1.1232 0.0503 0.9949
```

```
> exp((MuMIn::AICc(real.test.of.new.hyp) - MuMIn::AICc(prior.know1))/2)
[1] 7535.831
```

FIGURE 13

switch to a very mundane and non-linguistic example for a second, just because it is maybe easier to relate to in fact, and because there is a very nice data set in R available with which one can do that very easily. That example is concerned with how efficient are cars. We are going to look at the data here that measures this in this weird American way of measuring at miles per gallon, so that inversely related to what most people at least from Europe would be used to namely liters per one-hundred-kilometer. High mpg values means you can go a lot of miles with just one gallon of fuel. One gallon is about four liters and so obviously the further you can go with that one unit, you know, the more efficient the car is, so just bear that in mind. You are not going to need to interpret any statistics coming up later, but just so you know.

So if you were an engineer, then you might be interested in the question, what affects the efficiency of cars, if you measure it in mpg? So you do some exciting reading on this, and you think about it, and then you find a 1992 study that shows that the numbers of cylinder in your engine is related to that such that if you have a high-powered car with 8 cylinders or 12, it's going to be less efficient than the smaller one that has 3 or 4. Then you find another study from 1996 that says horsepower, so the performance of the car, is related to mpg, so again the stronger the car, the more powerful it is. A Ferrari goes less on a

gallon than, whatever, a small Toyota or something like that, because it has more horsepower.

Then of course you might just say, it's reasonable physics to assume that the heavier something is, the more energy it needs to be moved from A to B. I mean all other things being equal, the heavier car will be less efficient than the lighter car. Then, you come up with an idea, namely, there is another variable called displacement, so that's the engine size, for instance, measured in cubic centimeters or something would have an effect. Then you collect data, like I said those data are actually included if you install R, and you will find something very interesting, namely that just looking at engine size, so your hypothesis that the size of the engine has something to do with efficiency, leads to a super highly significant correlation: You can explain nearly three quarters of the data, of the mpg uses, just by looking at engine size, looks like total success. I mean you had this idea, you look at the data set, and you find that the effect is significant and it is really strong.

That's kind of what happens if you say 'frequency causes that'. You have some effect, in our case mpg, in another case it might be in reaction times or something, and you say it's caused by frequency—here, displacement—and then you do a monofactorial test that shows this is actually really strongly and significantly and highly correlated with that. So you have this and you write it up and send it off because how great is that? But, it's actually not great at all because this test is ridiculously anticonservative. What that means is you're testing your hypothesis that displacement is related to mpg against the null hypothesis that it is not, that's the logic of statistical testing: You formulate a hypothesis; you formulate the opposite, the null hypothesis, and then you run a test to decide which of the two is it. Here the output couldn't be clearer that it's your hypothesis. But what that ignores is everything else we already know. You're testing that displacement has an effect versus it doesn't have, but you're leaving out everything else we know has an effect, like weight, for instance, or everything else we know about this.

You're pretending we don't know anything about mpg, and I had this great idea and look at how significant it was. Statistically, the way to talk about this would be this: You're leaving up all the variability of the miles per gallon values, you're leaving all up for grabs for displacement. So there is a lot of variability in the data and you allow only your favorite predictor, displacement, to explain it. So obviously, whatever little bit of variability it can suck up like a Hoover, it will take, because you are not controlling for anything else. But that's delusional, because we *know* that cylinder and horsepower and weight affect mpg, that's our prior knowledge, that's what you knew before you started your



## What affects the efficiency of cars (measured in mpg)?

- In other words,
  - you're pretending we know nothing about **mpg** already
  - you're leaving all **mpg** variability up for grabs by **disp**
- but that's delusional/too generous: you already know that **cylinder**, **horsepower**, & **weight** affect **mpg**

```
> summary(prior.know1 <- lm(mpg ~ (cyl+hp+wt)^2, data=mtcars))
[...]  
Multiple R-squared:  0.895,    Adjusted R-squared:  0.8697 ***
```

- option 1: you need to test whether **disp** adds to what we already know

```
> summary(real.test.of.new.hyp <- lm(mpg ~ (cyl+hp+wt+disp)^2, data=mtcars))
Multiple R-squared:  0.896,    Adjusted R-squared:  0.8464 ***

> anova(prior.know1, real.test.of.new.hyp, test="F")
      Res.Df  RSS Df Sum of Sq    F Pr(>F)
2        21 117.16   4      1.1232 0.0503 0.9949

> exp((MuMIn::AICc(real.test.of.new.hyp) - MuMIn::AICc(prior.know1))/2)
[1] 7535.831
```

FIGURE 14

displacement hypothesis. If you look at these three things, look at how much they explain, not 71%, they explain nearly 90%. So 90% of the variable of interest, let's say displacement, let's say reaction times, is already accounted for by all those other things that we know.

So now you have two options. One is this: you need to test whether displacement, your new hypothesis, *adds to* what we already know. Again, in the frequency-reaction time example you want to explain the reaction times. You think it's frequency, but we know there is a ton of other thing that explains reaction times so what you shouldn't be doing is to check 'does frequency do something?', you should be testing 'does frequency do something given all these other things I already know about reaction times?' If we do this here, we get rather depressing news: If we add displacement, our favorite new hypothesis, to the model, it's adjusted *R*-squared value actually goes down. So your new hypothesis, your new favorite variable displacement, wasn't worth anything—in fact, it made the model worse than it was before. So conclusion, no, your hypothesis does not add anything to what we already know. Again, in the psycholinguistic example, that might mean, well, maybe frequency actually doesn't add anything to what we already know once you control for everything else that's around.

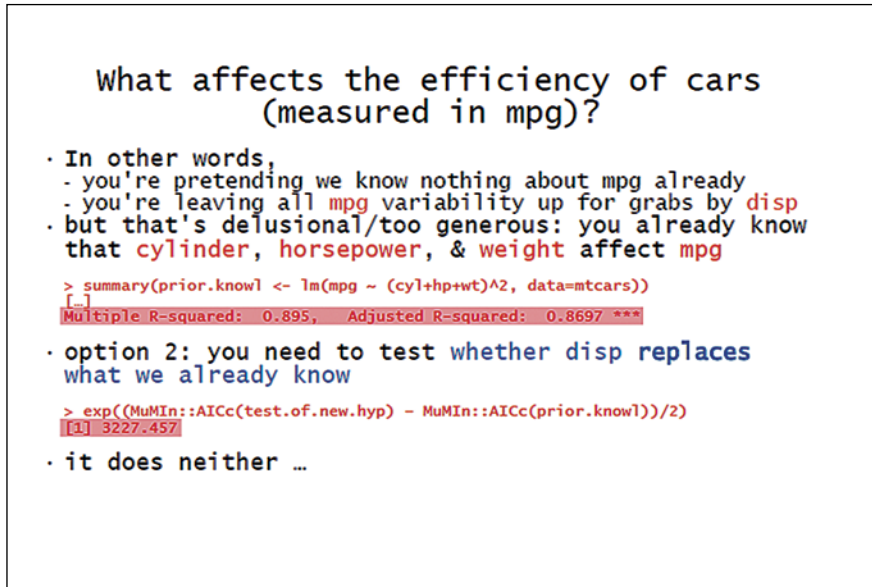


FIGURE 15

But there is a second option. You can either do this, test whether your thing *adds to* everything else, or you can check whether your thing *replaces* everything else. Maybe you came up with the one thing that does away with everything else. I mean not likely, I mean who of us is that great that they come up with something that disconfirms like the last 40 years' research? But it's possible.

Now if you do the relevant statistics here, this number, in a way that is not relevant right now, tells you that's extremely unlikely. Displacement does not do better at all than everything else we already know—in fact, everything else, what that number means is 'everything else, our prior knowledge, is 3000 times more likely to be the right explanation than displacement'. That's the technical way of saying 'you are wrong'. So it does not either. So basically, what we have here is a situation that shows that sometimes you might have a hypothesis about what a single predictor, displacement here, frequency here, might do, but in order to show that it actually does something as simple monofactorial test is not going to do anything—you need to keep and check everything else we already know about the phenomenon; then, things might look very different.

Another way of looking at this, somewhat polemically is if you look at this car data and you say displacement is important for mpg, then it's kind of like you do the following:

You look at this alternation here

- *John picked up the book.*
- *John picked the book up.*

So the question is whether people put the particle in front of the noun, noun phrase, or whether they put it after it. In this case, you can say both, it doesn't mean anything different. Then, we know that there is a length effect here: the length of the direct object has something to do with where the particle goes. The longer that thing becomes, the more likely it's going to be at the end. Because what you don't do is you don't have the verb and then you have a 20-word object and then you have a single particle dangling at the end. If the object is very long, you say *picked up ...*, and then you have the whole long object.

So again, hypothesizing that displacement is so important is to look at this, recognize that the number of morphemes is important, and then you have this great new idea that the number of syllables in the direct object might be a great predictor, when in fact of course that's going to be pretty much exactly the same as the number of the morphemes of the direct object. To the degree that we suffer from what Dąbrowska, what I quoted her from this morning, to the degree that predictors are collinear that weakens your case to make that a certain predictor is really all that powerful, if you don't consider all the other alternative explanations.

### But why would one do the latter, given all the empirical support?

- Because monofactorial views of the effect of frequency – any predictor – don't tell you much
- because correlations – between frequency & the linguistic phenomena – do not provide a clear enough picture of what's going on
  - correlation is not directional so the cause-effect relationship(s) is/are not always obvious
  - correlation is not causality: as above, frequency may be the cause, a cause, or just correlated w/ (a) cause(s) in a multifactorial setting
    - familiarity, concreteness, imageability, meaningfulness, age of acquisition (data from the MRC psycholinguistic database)
    - dispersion
    - length (another one of Zipf's laws)
    - other factors we'll talk about more later

FIGURE 16

### But why would one do the latter, given all the empirical support?

- Because monofactorial views of the effect of frequency – any predictor – don't tell you much
- because correlations – between frequency & the linguistic phenomena – do not provide a clear enough picture of what's going on
  - correlation is not directional so the cause-effect relationship(s) is/are not always obvious
  - correlation is not causality: as above, frequency may be the cause, a cause, or just correlated w/ (a) cause(s) in a multifactorial setting
  - familiarity, concreteness, imageability, meaningfulness, age of acquisition (data from the MRC psycholinguistic database)
  - dispersion
  - length (another one of Zipf's laws)
  - other factors we'll talk about more later

FIGURE 17

Now of course in the back of your head you're like, 'I mean, who the hell would do that? No one would be that stupid'. But that's not quite true. There is a lot of people who test their alternative hypothesis against 'we know nothing else' or 'we control for nothing else', that's not rare at all. Or, there is a lot of people that propose or compare models with such correlated predictors. That really happens a lot of times in general, [...] and I will show you here and how it happens with frequency.

Again, coming back to frequency here: The reason for you to maybe consider 'maybe frequency is not the greatest cause ever of these things' is because the monofactorial view of frequency or any other predictor doesn't tell you that much to begin with. That's because correlations are not precise enough to do exactly what we want. First, correlation is not directional, so from a high correlation, you don't know what's the cause and what's the effect—you only know that the two things are correlated. Second, correlation is not causality, and that is something everyone at least professes to know, that things can be correlated, but they might not be causally related. What might be these things that are related to frequency and might actually be responsible instead of frequency? Here is a few examples: for instance, familiarity and frequency are related: If you ask subjects to judge the familiarity of words, they will give you ratings that will be relatively highly correlated with their frequency. Same with concreteness, same with imageability, same with meaningfulness, all those were ratings

that collected from a variety of norming studies, for instances, in the 60's and 70's, and are freely available from the MRC Psycholinguistic Database and all of those are related to varying degrees to frequency.

Same with dispersion, the degree to which a word or a construction or a word in a construction is distributed evenly in the corpus. Another way of putting it would be: the degree to which, for instance, a word is used by every single speaker in the corpus and not just by three and the rest never uses it at all. That kind of notion of dispersion is also extremely highly correlated with [[frequency]] as you will see in a moment.

We all know that length and frequency are correlated. Things that are highly frequent in a language are usually short. If you look at English, *the*, *of*, *in* and *a*, those are most frequent four words in English, they are all supershort, and some other factors we'll talk about more later.

Let me show you some results here. Again, all based on the MRC psycholinguistic database. So in every one of these plots, you have logged frequency on the y axis. Frequency is based on the Brown Corpus of American English from 1960s, and again, these are logged. On every x axis, you have a different predictor or a different variable that is possibly related to frequency. Here we

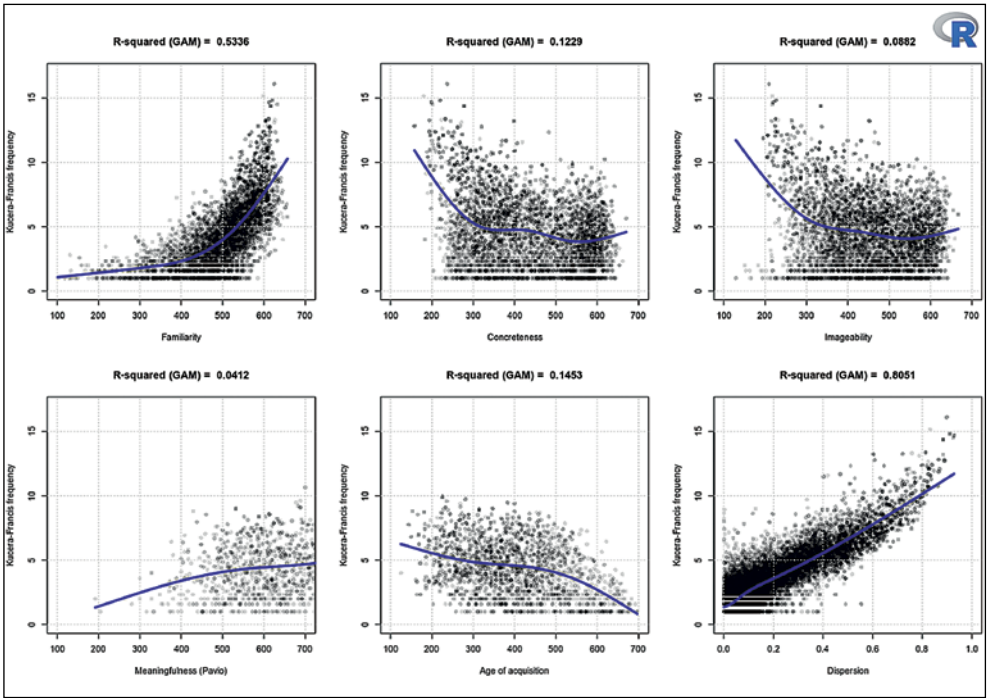


FIGURE 18

have familiarity scores from all these norming studies, concreteness ratings (so *apple* would score higher than *intelligence* or something like that), then we have imageability ratings (same thing imageability, *apple* would score highly because it's easy to imagine an apple in your mind, *intelligence*, it's difficult to imagine that as a mental image, something visual), then we have a meaningfulness scale here, same kind of task, age of acquisition, normed to fall in this range from 100 to 700. That's not days or something like that, and then we have dispersion, so the higher the dispersion value, the more evenly distributed things are.

Then here in these headings, you can always see the size of the correlation between things. You can see that dispersion and frequency are related by more than 80%. Frequency and familiarity more than 50%. Age of acquisition 15% roughly, and then the other three are a little bit less. The main point here is not to say 'toss frequency aside, use just this', but you can see that varying degrees of frequency can be explained by a whole bunch of other things. Just saying, 'it's frequency as opposed to dispersion' in a usage-based or cognitive context without ever looking at, 'but what are the relevant dispersion values in this language acquisition context corpus that I am looking at?' runs the risk of explaining whatever you are looking at with this axis (*y* axis) when you could just as well explain it with this (*x* axis). 80% of the data would be captured and I am pretty sure, and actually other people have done the work, if you take all these thing together and some other factors, there is nothing left for frequency to do.

If you subject frequency as a predictor to this kind of multifactorial way of thinking about things, then token frequency suddenly becomes much less important than many usage-based, but also psycholinguistic, works assume. A lot of other factors that are much less discussed in cognitive linguistic or usage-based literature are much more important. So that means frequency is basically 'overpowered' by other things. Other things just do better when it comes to explaining things, and these other things some of which you saw on the previous slide are correlated with frequency: obviously, you saw that on the graphs, but then here is the important part. These other things, for instance, or what some studies find, is these other things like familiarity or dispersion, various entropy-related measures we will talk about later, so those have an effect on reaction times, for instance, even if frequency is controlled for. So, you hold frequency constant in a typical statistical way and you check that something else, like familiarity, like dispersion still do something and the answer is *yes*. Then you try the opposite and you check, ok, if I hold dispersion constant or contextual distinctiveness constant, does frequency then still do something and answer is *no*.

## So what happens if we subject freq to a multifactorial way of thinking?

- There is an increasing number of studies that seem to indicate that at least token frequency is
  - less important than much of usage-based/psycholinguistic work assumes - other factors are (more) important, too
  - qualified/'overpowered' by other factors
    - with which it may be correlated, but
    - which have an effect if frequency is controlled for whereas frequency does not if other factors are controlled for
- many such factors can be gleaned from corpus data, but
  - their retrieval is less straightforward than frequencies
  - they are much less used - we don't do justice to
    - what the cognitive system does (Chater & Christiansen 2016)
    - Lakoff's (1991) cognitive commitment
  - some at least require - are you ready? statistical thinking ...



FIGURE 19

These studies show that if other things are controlled for, frequency doesn't have anything left to do. Frequency *only* does something if you don't control it with a variety of other predictors in psycholinguistic studies from, for instance, corpus data. The cool thing for a corpus linguist of course is that many of these factors you *can* get from corpus data, but their retrieval is less straightforward than frequencies which is one of those sad reasons why it does not happen often enough, especially not if you use web-based corpora. Many of the things you need to do in order to check whether frequency really has the kind of important effect that many people have been preaching, you can't do that with a web-based corpus—you need your corpus on your hard drive to do certain kind of searches and statistics that otherwise are not going to work.

That means of course they are much less used, and that means we actually don't do justice [[to]] what the cognitive system does, and what Lakoff's cognitive commitment says. If there is a psycholinguistic work that basically invites us, to put it mildly, to revisit the frequency effect, then maybe it would be a nice idea if in cognitive or usage-based linguistics we would at least entertain that thought and not just continue to pile on token frequency, token frequency all the time.



What are those other factors?

- Let me give you a fuller version of a quote you've seen before abridged)

Practice promotes proficiency (e.g., Anderson, 2009; Bartlett, [1932] 1967; Ebbinghaus, 1885). Learning, memory and perception are all affected by frequency, recency, and context of usage: The more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed (hence your reading this sentence more fluently than the preceding one). The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization; so a stimulus becomes associated with a context and we become more likely to perceive it in that context – all very rational.  
(Ellis, Römer, & O'Donnell 2016:45f.)

- much of what we all need to look at is in there
  - frequency, ok, but then recency (=priming & dispersion), context & conjunctions (=co-occurrence) ...

FIGURE 20

But that of course again requires of some imperialistic statistical thinking, which will raise a lot of people's neck hairs to a dangerous degree. What are those other factors that play a role? To answer that question, I am going to give you a fuller version of a quote before you only saw in abridged version for some very clever didactic motivation. Namely, the critical point now here, "learning, memory and perception are all affected by frequency [ok] but now also "recency, and context of usage: The more times we experience something, the stronger our memory for it [Ok, that's frequency. But then ...] the more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed". You can see here, "the more times we experience something", "the more recently we have experienced something", and then the effect is the same: "the stronger our memory for it and the more fluently it is accessed". They basically already make the point for us that, for instance, frequency and recency, they explain the same kind of things. Of course, they are related. If something is super infrequent, chances are you haven't seen it recently. But if you've seen something recently, chances are it's reasonably frequent unless there is a stronger topical or register or genre kind of effect that makes the occurrence of something more likely.



## What are those other factors & their effects on / relation to token freq?

- **Dispersion**
  - the degree to which X is distributed evenly in a corpus
  - strongly correlated with frequency (see above)
- **contextual distinctiveness** (Shillcock & McDonald '01)
  - the degree to which a word has an impact on the frequencies of the words around it
  - correlated with frequency, but outperforms absolute frequency as a predictor of lexical decision times
- **surprisal**: the degree to which sthg is unexpected
- **contingency** (Ellis 2006) (*can* be correlated w/ freq)
  - "it [is] contingency, not temporal pairing, that generated conditioned responding in classical conditioning"
  - "human learning is [...] perfectly calibrated with normative stat. measures of contingency like  $r$ ,  $\chi^2$  &  $\Delta P$ "
  - "[l]anguage learning can be viewed as a statistical process requiring the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations"

FIGURE 21

Finally, context of usage, "the more times we experience conjunctions of features", so context or the association between things, for instance, in the input or in what we produce is also an important source of how things might become associated with each other and how things might, for instance, speed up our psycholinguistic processing and/or behavior. What does that mean we need to look at? It means yes we should look at the frequency—I am not saying 'forget about it', I am saying 'contextualize it appropriately, given all other things we know that play a role'. Then these other things are things that I'll talk about during the remainder of this week. Tomorrow, I will talk about recency and what it does and how to measure it corpus linguistically and what we can do with that. The day after, we look at context and conjunctions of things, i.e. contingency or the association between words and words, or words and constructions, and other kinds of things like that. So just to give you some ideas ...

Dispersion, like I said, is the degree to which something is distributed evenly in the corpus. We've seen in this plot before that it is very strongly correlated with frequency. What I mean is this plot here: This was dispersion [back to Figure 18]. So again, the lower the value here, the more specialized words are. So here on the left, there are words that show up only once in a small corpus

part and never again. This word is *the*, the most frequent word in that corpus, and the second or third most widely dispersed one, I think in this corpus, ICE-GB data, this is like often, *and* or something. They are a little bit less frequent, but actually more widely used. So as you can see, frequency and dispersion, there is a big point cloud, but at the same time, *R-Squared* is in excess of 0.8. Definitely, a possible candidate for maybe being more important than frequency and we'll tomorrow see that it is.

Then there is this notion that I mentioned very briefly before, called contextual distinctiveness. That paper is an absolute stroke of genius. I think, it's not very much used in very many different contexts, but it's really insanely interesting. So it's a measure that does this: it quantifies the degree to which a word—I mean that is how they introduced it, it could be a construction as well—the degree to which a word has an impact on the frequencies of the words around it. To what degree does a certain word have a distinctive context, one that makes the frequencies of things around it be very different from what these things do in general.

So, for instance, one word might dramatically increase the probability that another word is used, even if that other word in general is very rare. So *hermetically* would be a case in point: After *hermetically*, there is only one word, namely *sealed*, *sealed* in general is pretty infrequent though, so there is a huge discrepancy between the frequency of one word in general and then what happens in a certain context. This is one of these studies that actually showed what I talked about earlier. This measure contextual distinctiveness is correlated with frequency: You would get a plot that shows a curve like this as well. But there are two interesting things about it: First, its computation doesn't actually involve absolute frequencies, so it's *related to* it, but it is not *based on* absolute frequencies in the simplest possible way. Then, more importantly is this. It outperforms frequency as a predictor of lexical decision times. So in this study, what they do show is that if you hold contextual distinctiveness constant, frequency doesn't do anything anymore. If you hold frequency constant, so that it can't do anything, then contextual distinctiveness *still* has an effect. Showing that this one is probably more likely to be the real cause of the effect, because it can do more things in more kinds of situations.

Then, there is this notion of surprisal that has been used in a lot of recent corpus-based psycholinguistic kind of work. What it quantifies is essentially the degree to which something is unexpected. It's *maybe* possible to talk about it or to consider it as an operationalization of salience, I am still not quite sure to what degree that works well myself, but I think it's *something* like that. I found some quotes in the literature that you will see on Day 4 that I think make that connection.

Then, of course, there is contingency or association, so the degree to which two things ‘like to occur together’ with each other, and that actually can also be correlated with frequency, depending on how you measure it. In other words, what that means is that people like Ellis (2006) would say, for instance, when you look at condition-learning, it is contingency, so the degree to which something is contingent on something else, the degree to which something is based on something else being present or haven’t happened before, “It is contingency, not just temporal pairing, or recency, that generates learning, for instance, in classical conditioning experiments.” The quote that you’ve seen before, “human learning is perfectly calibrated with statistical measures of contingency like  $r$ ,  $\chi^2$  and  $\Delta P$ ” and others like that. Then, finally, “language learning can be viewed as a statistical process requiring the learner to acquire a set of likelihood-weighted associations between constructions and their functional/semantic interpretations”, which of course, I mean, especially after the introduction this morning, constructions are form-meaning or form-function pairings, so obviously acquiring when and when not to use a certain construction makes it necessary for you to know, ‘this thing can mean that thing, but it can’t mean this thing unless something else changes in the linguistic configuration or the linguistic context’.

So as Nick and others wrote, “it is not enough to ‘know’ a construction in the sense of entry in a dictionary or grammar book. [And as I would insert here, ‘and it’s not enough to know just how frequently does it happen’] You need to

So, it is not enough to "know" a  
construction in the sense of entry  
in a dictionary or grammar book.  
You also need to be able to bring  
the appropriate interpretation to mind,  
and that involves knowledge of  
its association strengths.

Ellis, Römer, & O'Donnell  
(2016:36f.)

FIGURE 22

## What are those other factors & their effects on / relation to type freq?

- **Type-token ratio**
  - the ratio of types per token (quite dependent on text/corpus length)
  - alternative measures esp. from LCR/SLA
    - repetitiveness measures such as Yule's  $K/I$  or Guiraud
- **entropy  $H$**   $H = -\sum_i p(x_i) \log_2 p(x_i)$  ( $w/\log_2 0 = 0$ )
  - the evenness of the type-token frequency distribution, which can distinguish distributions with
    - the same numbers of tokens
    - the same numbers of types
    - and thus the same type-token ratios
  - the more extremely Zipfian, the lower the entropy
- nearly all of these will be discussed later to give you a good understanding of what corpus data really have to offer

FIGURE 23

be able to bring the appropriate interpretation to mind [so what's the function component of it?], and that involves knowledge of association strengths" (Ellis, Römer, & O'Donnell 2016:36f). That's something we will talk about a lot on Day 3.

What else is there? Obviously, we looked at type frequencies and at token frequencies. So much of your research should at least be taking care of on some level, type-token ratios, how many different types you see per individual token. The problem there a lot of times is that this is the measure that's very dependent on text or corpus length. So alternative measures, in particular, maybe from second language acquisition research or things like that, would be more useful. For instance, there is a bunch of relatively nice—and easy to compute, actually—repetitiveness measures. In a second language acquisition context, what those would indicate is to what degree do learners of English, non-native speakers, always overuse the same words as opposed to having some lexical variety. Obviously, the idea would be 'the better your command of a foreign language, the more non-repetitive you are'. Those things would be interesting, because in the construction grammar or usage-based kind of corpus context, it would mean that you use not always the same words with the same construction—you know it's a construction because you are able to put many different things into its relevant slots.

## What are those other factors & their effects on / relation to type freq?

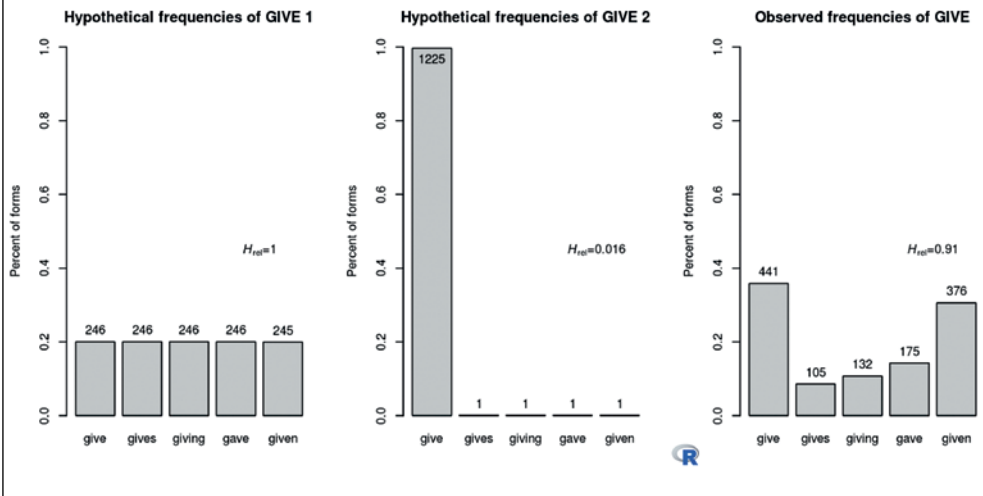


FIGURE 24

Then something we'll talk about a bit on Day 4 is this notion of entropy, and entropy is quite interesting and in cognitive linguistics at least I think completely underutilized, although it's very useful. It's concerned with the evenness of a type-token frequency distribution. What that means is you can compute this measure to distinguish distributions that have the same numbers of tokens and the same numbers of types. Obviously, if two constructions, for instance, have the same numbers of tokens and the same numbers of types, the type-token ratios would be the same because you'd be computing it on the same numbers, but that doesn't mean that the distributions are actually the same.

I'll show you very briefly here a graph that highlights this. These are some data, some of which are made up, some of which are real: The frequencies of the forms of the verb to *give*. Here is the distribution as you can see that has been created, to be as equal as possible: *Give*, *gives*, *giving*, and *gave*, they all occur 246 times, and one of them (*given*) occurs 245 times because the sum of this is what actually happens in that corpus of 1,229. But the point is this is 5 types, 5 verb forms, the overall frequency of them is 1,229, but you can have the same number of types and the same number of tokens and the same type-token ratio with *this* distribution: Again, you have five types. You have 1,229 tokens, but obviously, one of them is nearly everything and then the rest hardly ever exists.

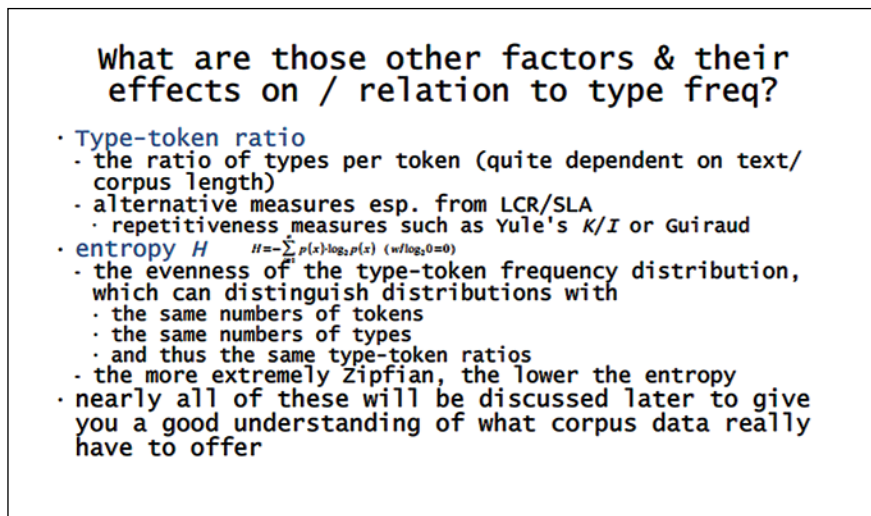


FIGURE 25

And then this is the real one. In this corpus, *give* was the most frequent verb form followed by *given* and then the others are much less. All three distributions have the same type frequency, the same token frequency, the same token-type ratio, but they differ dramatically in their entropy.

The entropy values I've computed here range from 0 to 1, so this is close to the theoretical minimum, this is, with rounding, at the theoretical maximum, and the actual distribution is much closer to something like this than something like this. So relatively even distribution compared to this. If you want to talk about what happens in constructional slots, how many alternative things can go in there, and how are they distributed, just the type frequency isn't going to do it. I mean, maybe you can do *something* with it, because type-token distributions on the whole will be Zipfian, but to get a better understanding of what happens, just type frequency, just token frequency is just not going to cut it, you need something more precise and entropy is a very simple way to compute something like that.

In fact, this entropy measure *can* be related to the 'Zipfianness' of the distribution. The more extremely Zipfian [becomes], the more the curve doesn't look like this, but like super steep and then leveling off, the lower the entropy value will be. So anything having to do with learning/acquisition of categories *will* be related to entropy because, for instance, children will need to see certain path-breaking verbs, for instance, in a construction enough times, but then also there needs to be a tail of other things so that the kid figures out actually any verb can go in there with a certain semantics or something like that. As

## So what does (token) frequency really do, and does it matter?

- On top of some other studies I will mention later, Baayen (2010:436) is one of the most comprehensive studies of the word frequency effect
  - after controlling for many local synt & morph. cues
    - dispersion, contextual distinctiveness, entropy, many more ...
  - he finds that
    - "[f]requency emerges as the best single predictor, as expected", BUT
    - "frequency of occurrence, in the sense of pure repetition, turns out not to be a particularly important predictor"
    - "frequency-as-repetition explains only a small proportion of the variance in response latencies"
    - freq-as-rep - exactly what was supposed to be sooo important
  - so, no, not everything is explainable w/ frequency
  - and for the scared anti-imperialists: this is not just number-crunching - it's theoretically important
    - if freq-as-rep is a causal factor, then maybe entrenchment / resting-activation-level accounts are correct
    - but if it's not, we need different theories/models ...

FIGURE 26

I already indicated, most of these things will be discussed in the remainder of this week to give you an idea on how you can use them in your own research.

To wrap up, what does token frequency really do and does it matter? Well, it depends a little bit on what you think the workload, so to speak, the conceptual workload of token frequency is supposed to be. One of the coolest study in this area is Baayen (2010), who looked at that the word frequency effect, in fact, the title of the paper is, *Demythologizing the word frequency effect*. What he did is, he did a really really comprehensive corpus-based analysis and in that analysis, he did what I talked earlier about this car example: He *did* measure all these other things and he did control for all the statistically to *then* see is there actually anything left for frequency to do? He controlled for what I call here local syntactic and morphological cues—some of those things were like dispersion (how evenly are things distributed in the corpus?) contextual distinctiveness (how much of the impact does a word have on what happens around it?), syntactic and morphological entropy (how many constructions does a word show up in, how many different constructions does a word show up in, how many different morphological forms is a word used in, and what is their distribution?—so he measured all these kinds of things.

What he finds is this, “frequency is the best single predictor [as expected *but*] frequency of occurrence in the sense of pure repetition turns out not to be a particularly important predictor”. In fact, I think in the final analysis that he

reports, it's not even significant any more. I mean, I am not even sure whether that sinks in so quickly—of course it's just one study, but if that were to hold up, I mean that's a very interesting result because it's essentially the kiss of death to every account of entrenchment and everything in terms of a "repetition counter" and "resting level of activation gets raised" and something like that. Because "frequency-as-repetition" counter, if that were to hold up in subsequent studies, doesn't do anything.

Another quote, "frequency-as-repetition"—in exactly the way that the Langacker quotes, the Schmid quotes, and to some extent the Ellis quotes, in exactly the way that they say frequency is doing all this stuff—he finds that, for instance, response latencies are not accounted for very well. Now you might say, but maybe it has these other factors in cognitive and usage-based linguistics, and maybe it does, but again remember, for instance, the Schmid quote: it was specifically about "speed and ease of lexical access", also known as "response latencies", so specifically, the type of effect that frequency was supposed to have in Schmid, who is one of the leading people on entrenchment, Harald has shown it doesn't do anything there. So exactly the type of thing that was supposed to be so important for theory formation ... not so much. So no, not everything is explainable with frequency is one of those things that that shows.

Second, that is an important point to make, given some of the sentiments we've seen earlier today, for the sort of scared anti-empirical imperialists, this is not just some anal number-crunching ('Let's put in some another thing, and then we find something else')—I mean it's theoretically important, because like I said, if this were to hold up, then frequency-as-repetition explanations in terms of a resting-activation-level counter or the kinds of things that in cognitive linguistics are routinely used are just wrong. There is a point in which the proper kind of empirical analysis does inform theory formation if only by ruling out, for instance, the important role that a certain theory has attributed to a certain predictor, here, frequency. If frequency-as-repetition is not important, we need different theories or models to explain the kind of effects that we are concerned about so far, because the above quotes by Langacker, by Schmid, and to some extent by Ellis, they are actually not explaining things right, right? I will talk about this a little bit later on Day 4, but basically if Harald's analysis is the right one, we need a different kind of psycholinguistic model, namely, one that has more to do with 'how evenly and how widely connected are nodes in our linguistic network?' for lack of a better term as opposed to 'how frequently is something called and what is its activation-level?'. That's a totally different kind of psycholinguistic model, but it's the one you only get to, if you do the serious kind of analysis that he has done to show that frequency as a repetition effect actually does not do much. Thanks.



## Frequency: Practice with R

Today's session will be tricky, I have to say, because R is a very powerful tool, it's a very powerful programming language, but that of course also means there's a little bit of a limit on how much one can cover in any one set of short sessions. What I'll try is I'll give you at least a good overview of the kinds of things you can do and I'll discuss using some code examples that have been made available, some applications of how you can get very simple things done. This kind of practice sessions are going to be extremely different from the talks: There's not going to be a lot of slides or anything, not going to be any theoretical stuff, but rather it's going to be about data processing with R of a certain bunch of corpus data.

The first thing you will need in order to follow along is the data. If you look up my name in a search engine of your choice and you go to this website, then if you click on my name here, you can download a zip file. Those of you who want to follow along now, please do that now. Just click on that, download that zip file somewhere.

Once you have that file, you need to please unzip it, so, extract it. Right-click on it and then "extract all" or "unzip here" or whatever it says in your operating system. It is important that you unzip it, you cannot just leave it there as a zip file. You need to make it a regular folder and if you unzip it, it should look like this. So what you need to have is a folder that contains these files, the "03" data file here in particular, and then these four code files. This folder contains these five things. Four of these things are R scripts, scripts that do certain different types of corpus analysis. Today we're looking at the `03_frequency practice`. Then, on the other occasions, we'll look at dispersion, association, and the last one, `practice.r`. The folder `03_data` contains some corpus data. Specifically, as you can see here, it contains two different corpora and one other text file.

The first corpus it contains is the Brown Corpus. That's a corpus of written American English from the 1960s, relatively small by today's standards, just



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](https://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611276>

one million words. The second folder here, `ICEGB_sampled` contains all the words from the ICE-GB corpus, but it doesn't contain the original format of the ICE-GB corpus because I don't think I'm allowed to share that. But it contains all the words that the ICE-GB corpus contains with their specific annotation. Then the last file is one of the files from the Brown corpus, but not tagged. This file is available here as a tagged version and here available as an untagged version, so that we have different things to play around with as we proceed.

I don't know how many of you have R and RStudio installed. If you do have it installed, then one thing you could do now is open this `03_file` within RStudio and then that will look approximately something like this. This is the code file we'll be working with today. If you do *not* have R or RStudio installed, that's fine. You can follow along anyway, in that case just open that file, that `.r` file, with a text editor. If you're using Windows, then that might be something crappy like Notepad; if you have something better installed, like Notepad++, that would be more useful, but any text editor will be able to open that file. This is what you should see in that file when you open it, the stuff here on the top left is what you should be seeing. I'll give you one more minute or so to get ready. Then I'll start discussing this, i.e. what this code does. [Prof. Gries interacts with the audience for few minutes]

Let's get started. R is an open source programming language. It is, strictly speaking, a general-purpose programming language, but it is one that has been designed for statistical analysis in particular. But again, it's a general-purpose programming language, you can do all sorts of things with it that have nothing to do with statistics. If you wanted to write a backup software with it, you could do this. If you wanted to write a web crawler with it, you can do this—I have. These kinds of things are all possible. But again, the design is such that it makes it very easy to do certain types of statistical analysis, actually, pretty much all kinds of statistical analysis. At the same time, it also allows you to do all sorts of text processing with it as well. A lot of people have been using a Perl in the past, many people are still, of course, using Python for text processing or any kinds of natural language data processing, I'm using R all the time, the main reason for that being that, first, I think it's simpler than definitely Perl, maybe than Python. Secondly, I like the idea of being able to use one and the same programming language both for everything that has to do with my text processing and then also for the statistical analysis. Another way of putting this would be, I'm happy that I can be lazy. I don't need to learn different tools. I can do it all within one.

The way R works is actually relatively straightforward if you've ever worked with a spreadsheet software. For instance, here is a spreadsheet software that is relatively similar to Excel. The way that spreadsheet software works is by using

functions. For instance, if you have some data in these four cells, and then you want to compute what the sum of these four values is, then what you write in the cell below is equals, and then you use that function, SUM, which says, 'compute the sum of something', and then you tell the function to compute the sum of what, the sum of these four cells here, then closing parenthesis. Then once I press Enter, then we'll have the result ten here. R basically works exactly the same way: You have functions that tell R what to do and you have things in parentheses, so called arguments that tell you, for instance, what the function is applied to. This `[[=SUM(B2:B5)]]` means apply the function SUM to the contents of the cells B2 to B5, same with if I had written AVERAGE here or something like that. That's basically exactly the same way that R works just that it's a general-purpose programming language so there're many more functions than the spreadsheet software which can only handle certain types of data and certain types of input.

What I want to do, basically, is walk you through the code that you have in the program book, which is an output file from what happens when you run this R code and then basically explain to you a little bit about how the code does what it does. But I have to say again, there's a limit on how much in detail I can go here: You don't learn a programming language in four sessions over the course of a week. What you can do is, you can get an initial understanding of how these things work, but then if you're convinced, as I think you should be, that this kind of stuff can help you with corpus-linguistic analysis, then obviously, you will have to follow up on that. It's probably not going to be a big secret or a super shameless plug to say that I wrote a textbook on how to do this kind of stuff,<sup>1</sup> basically explaining how all sorts of corpus analysis can be done with an R, even things that probably go way beyond what you're thinking is possible. But again, we'll only scratch the surface here, but it'll give you an initial idea of how things work and then, hopefully, you'll be able to follow up from there.

My understanding is you have this in the program handbook. What you can do is, you can follow along on this. If you don't want to run this stuff at the same time, let me just explain to you what this looks like. Everything that has gray shade here behind it, like this, that's programming code. That's stuff that you also have in that `.r` file in that little script.

Then everything that has a white background, like here, and has these pound signs at the beginning, that's output.

So basically, this does something, then this generates some output, and that's that output. Then here's the next line of code. When you scroll down or,

---

<sup>1</sup> *Quantitative Corpus Linguistics with R*, 2nd revised edition, Gries, 2016.

in your handbook, turn the page, then this is some code that produces that output, and so on. So you can always follow along and see what is it that a certain line of code or certain lines of code produce to, later, maybe follow up on what is happening here.

The other thing I want to draw your attention to is this. Now let me show this in the html that's probably better.

What I've done here all the time is I've tried to make this code as comprehensible as possible. This is the R code that is actually doing something. `corpus.file <- scan (blah ... blah ...)` we'll discuss all this in a moment. Then, everything after pound signs (`#`) is commentary that I put in there to explain to you what that certain line does. There is a function in R called `scan` and just like in the spreadsheet software, there's a function called `sum`. So that function `scan` loads files. I'm saying here: `scan` does some loading and then `scan` takes four arguments. One, two, three, four, each is on a separate line and the reason for that is that at the end of every one of these lines, I can tell you what it does. So basically, what I've tried to do is, after the comments, always provide you English translations, so to speak, of what the R code means and what it does. Again, once you follow up on this, basically if you read the dark part here, so the black font and things like this, this is the programming language, and then this is the English paraphrase or translation of what that programming language does and what each of these arguments does. Especially, if you want to add notes, then maybe first look at, maybe what you want to write down is already provided here as commentary to help you understand better what the text on the left, the code, is actually doing.

Let me use this to walk you through this. The first two lines are relatively cryptic. I'm not going to talk about them much. The first line just makes sure that the R work space is empty so, basically, that would be the equivalent to say that in [[Microsoft]] Word, no file is loaded, in a spreadsheet software, no data are in the spreadsheet—just clears the memory, so that there's nothing left from a previous session that messes with your current analysis.

If you do a lot of work in R, obviously, it might mean that you have worked on different projects and so what this first line does is, it makes sure that there's no data in R still from an earlier project that messes up maybe what you're doing in your current analysis. The second line here loads a particular library into R so that whatever that library offers, in terms of data and functionality, can be used. So the thing is, like I said, R is an open source programming language. What a lot of people have done across the world is they've basically written little scripts and packages that augment, that add to the functionality of R, allowing R to do things that before were not possible. So right now, I think that's probably like ten or twelve thousand packages out there, where people

have written these things so that other people can do the kind of analysis they have done and especially that extendability is one of the reasons why R has become so popular. What these days happens is that if someone develops a new statistical method, you can be pretty sure that the first implementation of that method will be available in R, because then everyone in the world can download it ten minutes later, test it, run it on their own data to see whether it works and things like that. Same with problems in R. If there's ever a problem in R, sometimes, I've had the case once, you write an email to someone who maintains the language, and like a day later, there's a fix available online for free. If you compare that to commercial software, sometimes it took years for commercial software to fix errors. In this open source environment, it goes much faster than that.

What we're going to do here in this part is we're going to work on frequency lists, because that was the topic of yesterday's talk. We're going to look at different scenarios. The first and simplest one will be to create a frequency list of a single file. That's typically probably not particularly useful, because most corpora come in many different files, potentially thousands or tens of thousands of files. But we have to start somewhere, and the easiest way to start is with a single file.

The first step, obviously, if you want to create a frequency list of a file, is you have to load it, duh, not a big secret there. In order to load the file, we will need some things. First, we need this function `scan`. `scan` is a function that, among other things, can load files. Actually, it can also load websites, for instance. That's how you would, maybe, write a web crawler or something like that. We need this function `scan`. That function `scan` then takes parentheses and four different arguments, as you can see. Then, what we need to do is when this scanning processes, scanning a file, all this stuff which I'll discuss in a moment, when that is done, then R picks up something from the hard drive, loads it, and then what happens is this: this little syntax thing here `<-`, it looks like an arrow, right?, like a left-pointing arrow, so that means, 'the result of this, of the scanning, put it into an object called `corpus.file`'. That's why a paraphrase is this here, 'make `corpus.file` the result of' that is this: `[[corpus.file]]`, and then the arrow `<-`. You can read the arrow `<-` something like 'is supposed to be' or 'is supposed to become', `corpus.file` is supposed to become the result of scanning this file in this way. Once R is here, at the end of this, then the contents of this file will be in an object called `corpus.file`. Nearly all of the stuff R works in that way, you create something and you save the result of what you created in a so-called object. That object has a name, you can call it whatever you want—I mean with some limitations, but theoretically, whatever you want—if you wanted to, you could call it Peter or Fred. Not that that would be

useful, but whatever. So here I'm calling it `corpus.file` because that's what it is. What is `corpus.file` supposed to be? It's supposed to be the result of loading. The first argument to a function that loads the file should be the file that is supposed to be loaded. This refers to the file `Brown1_J_untagged.TXT` in this folder, all the data that you downloaded from my website. If you're in the directory that contains this folder, then this line can read that text file. Now, if you load text files, there are a few other things you have to tell R what to do with it because R is a statistical programming language at first, so it's mostly used to loading numbers—because you do statistics on numbers.

If you load something that is text like this, you have to tell R that. So that's what the second argument does here: `what`, the argument is called `what`. You're telling R what it is that you're loading. You're telling it that you're loading character strings [`what=character()`]. You're telling R also that the character strings are separated by line breaks, that's how you write it like this [`sep="\n"`]. Here, I'm suppressing output that says how many lines were loaded. Again, `corpus.file` is supposed to become, or be, the result of scanning this file, which contains text, or character strings, that are separated by line breaks, and do the loading quietly, don't give me any output. At the end of running this, that loading process has happened. Let me show you what that will look like in R: So, `corpus.file` is supposed to be the result of loading this file, which is character strings separated by line breaks, don't give me any output and I was in the wrong folder.... Let me run this again from the right directory. Nothing seems to have happened, it just concluded, but that's actually good thing, because we said, we don't want to see any output. If we had seen any output, that would probably be an error message or a warning or something so the fact that we didn't get any of that is actually good news. So now we have an object here, `corpus.file` that contains the contents of this file from the Brown Corpus. Because we just loaded this `corpus.file`, in this environment tab here, it now says, this object exists: `corpus.file`, it's a large character vector that actually contains many ten thousands of words. We'll see how many words in a moment. So the first thing you want to do, whenever you load a `corpus.file`, is you want to make sure that the loading was successful. So that's what I'm doing here. There's a function `head` that shows you the first N, whatever number you enter, the first N elements of what you just created. So here I'm saying, show me the beginning of the `corpus.file`, the first ten lines. That's what you see here: you see the importing of the text file has been successful. Now, `Brown1_J_untagged`, that corpus files is in R for further processing.

There's a similar function called `tail`, which shows you the bottom of something. `tail` shows me the bottom of `corpus.file`, namely the last ten lines. Here you can see, this is the end of that corpus file, the last ten lines that we've

just loaded. It's always a good idea if you work—I mean with any data, actually, not just text data, but also statistical data—it's always a good idea to make sure that the importing of the data into R has been successful so that you know, I can continue with writing code from now on. It looks pretty straightforward.

Now, if we want to create a frequency list, then, obviously, we have to prepare the data a little bit, because for one thing, you can see that every single element here, every single line is not a whole sentence, but it's also not a word. Obviously, there are many words here in one line. If we want to count the words, we need to cut up the lines wherever there's a new word, so that we can count those. Right now, we don't have every word on its own, so we can count it—we have parts of sentences, namely, roughly eighty characters of sentences from this old legacy format. That's one thing we need to take care of.

Another thing is that we would need to consider [that] you can see all sorts of annotation stuff that we don't want. So probably you don't want to count this, things like this `[-\|q]`, all sorts of things in there that maybe you want to get rid of. Probably they're not going to hurt much if you leave them in, but they're going to clutter up the results. The more precise you need to be, the more you would want to get rid of those. The biggest thing you probably want to do is to get rid of this line by line annotation here at the beginning. This is, the first letter here, is the part of the Brown corpus, namely J [referring to the file `Brown1_J_untagged`], that's the file we just loaded. Then there's the file, and the component in J [referring to the file `Brown1_J_untagged`] and then there's a line number. Obviously, you don't want to count those. The first thing we're going to do is actually, we're going to clean up the beginning of the lines here in order to make sure that when we do the counts, we don't also count this kind of stuff. I mean, what do we care how often “J80” shows up in corpus file or something like that.

So the first thing then we'll want to do is to use a function called `gsub` to clean up the file. What the function `gsub` does is basically, it's implicit in the name, the “g” stands for ‘global’ and the “sub” for ‘substitution’. So basically, “global substitution”: ‘replace everywhere’ is what that means. The function `gsub` takes minimally three arguments and most of the time, it's really relatively intuitive even if it doesn't look like that, it's relatively intuitive. What they are—if you want to replace something, then obviously you have to say what do you want to replace, by what do you want to replace it (you know, what should go in there instead), and then what object do you want to apply this to: I mean, where is the stuff that you want to replace, kind of makes sense.

What we're doing here is we have this object called `corpus.file`, which contains all this. Now we're creating a new version of `corpus.file`: Make the new version of `corpus.file` be what you get when you replace this (I'll explain this in a moment), this `[“ ”]` by nothing: these are two double quotes with

nothing in between. If you replace something by nothing, then you're deleting it. Do whatever deletion this refers to, do that in `corpus.fi1e`, and then there's one additional argument: Whenever you want to do text processing stuff in R, you probably want to say, you're using Perl-compatible regular expressions. Perl is one of the most powerful programming languages, or has been in the past at least for text processing and so we're using basically the capabilities that Perl offers as a programming language, we're using that in R here.

Now, what are we replacing? That's this `[[gsub]]`. You see here already `[[“^.....”]]` what it means, the first nine characters of each line. Where are the nine characters? Well, the line starts here with a “J”, so it's one, two, three, four, five, six, seven, eight, nine. So that's what we're taking out because it's stuff we don't need. So again, we don't have time to go through a whole discussion about regular expressions here—I mean, that would be a workshop on its own—so I'll always explain to you a little bit what the regular expressions do but in order to write them yourself, you will have to read up on this in some textbook, maybe mine, maybe others. So what this syntax means is, this little roof thing here, the hat sign or the caret ^, means ‘at the beginning of a line’, that's what we have here: Every one of these things we want to get rid of is at the beginning of the line, and then a dot `[.]` just means ‘any character, anything’. So the dot will always be the “J” and then sometimes the eight, then the zero, and then the space so this is like a placeholder or like a wild card that you might have used in Word or something like that when you say “find anything like this”. So here, then, I'm saying ‘find at the beginning of the line, one, two, three, four, five, six, seven, eight, nine characters, and replace them by nothing’. If we do that, it looks exactly like what we want, you can just about see it here. Look here at the top. Here's the fourth example, fourth element of the `tail`, which is *J80 1830 earth. The point is*, and look here is the same thing without the beginning: *earth. The point is*, now we've taken out what was in front of this before the line, the corpus identification thing has gone, and the line number is gone. Now it already looks a lot better, because we don't have to deal with this line annotation stuff. Now things look like this. Now the next thing you want to do is, generally what you would need to do is split this thing up into words so that we can count them. Because here everything is split up into roughly eighty-character sequences that contain words, but also other things—we want to have every word on its own.

The first thing we will want to do, pretty much always, if you especially work with the Latin alphabet or something like that, as you want to get an inventory of all characters that are attested in your file, so every single letter, number, punctuation mark, every single character that's in the file you want to see, and the reason for that is mostly sort of a being-pedantic-while-you're-programming kind of reason: You want to know, if you work on a language



like English or something, you have to make decisions about what is a word. Obviously, you can think at least that everything that's surrounded by spaces is a word. But then what about things that are hyphenated? If you have *well-behaved*, is that one word or two? Or if you have *co-occurrence*, is that one word or is that two? You need to decide what to do with hyphens. But of course only if there are hyphens in your data. So we're doing this to figure out, do we need to care about hyphens or not? Same with apostrophes: In an English text, there will be apostrophes, but in other languages, you might not have apostrophes much. For instance, in English, do you want to treat an apostrophe *s* as part of a word or not? *Peter's car* is that two words or three words? That's something you need to decide if you write a program that counts words. Of course, in English it is super annoying because the apostrophe *s* can be the genitive *s* or it can be the third person singular *s*, or even the third person singular *has*. That would be something you need to take care of.

Second example, for instance, what do you do with numbers? If you have *15-year old*, then it's *15-year*. Do you want to treat the *15* as a word, or are you saying, 'I don't care, whatever, out with it'? That kind of stuff is the reason for getting a character inventory, so that you know what you need to deal with potentially. Same with dashes, other kinds of punctuation marks, and things like that.

Punctuation marks can be really tricky in English. Normally, you would think there's a period at the end of a sentence so that's what we can split up. Well, but *Ph.D.*—Ph dot D dot, so abbreviations also have those so you need to deal with that. All these kinds of things are the reason why I'm doing this here just so that we get an idea of what are all the characters that are listed in here.

Now, just to make that very clear, a lot of people, when I talk about this stuff, you can see their eyes glazing over, because they're like, 'OMG, I have to deal with this kind of crap. I'm never going to touch R again'. Well, but that's premature because actually, if you work with any other corpus software, you also have to deal with that. The only difference is that other corpus software hides what it does in some opaque settings or initialization file, and maybe doesn't even show you what it does but the decisions are made. The question is whether *you* make the decisions because you look at this or whether you trust that Wordsmith or AntConc or whatever software you are using makes the right decisions for your data. The only difference here is that this makes it explicit, whereas, on other occasions, you might just trust that a software does `[[make]]` the right decision. That can be really wrong. As I've discussed in one case in my book, there was one case where—I'm not going to say which—but one software was upgraded over the web so you could download a patch in it, updated some sort of things, and it made some changes in terms of how certain

things were computed. But you didn't have a file that also said what exactly the change was. What could happen is, you do a certain analysis on Monday, just trusting the program, then you download the upgrade on Wednesday, and then you rerun the analysis on Friday, get a different result, and don't even know why. I'm not even saying the improvement was a bad thing, but what I'm saying is, if you do it like this, then at first it's a pain, because you have to be so freaking explicit—but at the same time, at least you know what you're doing and you don't trust any person halfway across the world that they did the right thing for your data.

What we're doing here, then just very briefly, is there is a function called `strsplit`. What it takes is it takes text and splits it up by something and the something here is an empty string: this is two double quotes again with nothing in between. If you use that with `strsplit`, what it does is it takes the text and splits it up character by character, every single character. Then this opaque symbol here, that's a % and a > and another %. What that does is it says 'take the result and use it in the next thing'. So this says 'split the `corpus.file` up character by character, when you're done, give the result to the function `unlist`, which turns it into a vector, if you're done with that, give the result to a function called `unique` [`unique` does something which is very easy to explain to linguists, namely, it creates a list of types out of a list of tokens so it looks at everything that's in there attested at least once], when you're done with finding the types, then take the result and hand it to the function `sort` so that it gets sorted alphabetically', and then we get this.

So now you see every character that's in that Brown corpus file, space, underscore, I guess, a hyphen, comma, semicolon, and so on, so a whole bunch of non-letter stuff. Then here the letters begin, "a", "A", "b", "B", and so on. Now, for instance, we see that actually, if you work with this, you need to decide what to do with percentage signs, you need to decide what to do with the ampersand, with a smart apostrophe like that, the tilde, all sorts of things that you would need to deal with in some way.

Now we're going to do it a little bit easier here and use a shortcut that is very convenient, but it's very similar to what I've showed you. So we're doing this: I'm going to skip over the regex first to show you the whole thing and then I'll discuss the regex. So split up the `corpus.file`, which still looks like this, so split up that `corpus.file` by something—I'll tell you in a moment—using Perl-compatible regular expressions. Then, when you're done with splitting, hand the result to `unlist` so we get a character vector. If you're done with that, hand the result over to the function `tolower`, which converts everything to lowercase—for this example here, we don't want to distinguish between upper and lower case, we're just going to use all lowercase for everything—then,

when you're done with this, put the result into a vector called `corpus.words`. So, split up by something, make that a vector, make that small letters, put it into `corpus.words`. That's how we're getting the words. Now what is the regular expression we're using for splitting here? Again, we can't even come close to discussing regular expressions here in detail; what this means is basically 'find everything that's not a letter'. The plus means 'one or more times'. So that means, for instance, R goes through the `corpus.file` like this and says 'here's three letters, and then there's a non-letter—oh, ok, this is where I'm supposed to split this up', so this is how the *the* will be isolated. Then it looks at the "i", "n", "p", "u", "t", and then there's another space, so here's another thing where it gets split up. Then here's another space where it gets split up. So it goes through it line by line, character by character, and all the time checks "Is this a letter, yes or no?" If it's not a letter, then, there, it splits it up. Same here: It sees the five letters of the word *level* and then it sees a comma and a space, and a tilde and a pipe, all those four are not letters, it's one or more non-letters. So these get deleted by R splitting it up. So it has *level*, and then it has the *e*, and then after the *e* there's a space so then this gets eliminated. Then there's the *is* and so on. So line by line, character by character, it checks, 'what is that?' If it fits the description, it gets deleted and split up.

Then we have an object called `corpus.words`. Then it's actually really simple. This is something that I like about R a lot: Sometimes its syntax and its functions really look a lot like plain English. So if you want a frequency list, you just say, 'create a table of the words in the corpus and when you have that table, sort it in decreasing order'. Then we look at the first thirty items of this, and this is what we're getting. As usual, *the* is the most frequent word, and then *of*, *and*, *in* and so on. Those are always among the top six or ten that we find.

That would be the simplest way to get through the steps of creating a sorted frequency list. As you can see, it's sorted by frequency because we said `decreasing=TRUE`. The highest frequencies come first, and then the numbers get smaller and smaller, which means at the bottom, you will have all the words that occur just one time in this file.

Like I told you before, a Zipfian distribution, there is going to be very, very many words that occur very little. In fact, we can plot very easily a graph that shows this Zipfian curve here. I'm plotting the typical kind of Zipfian curve, namely on the x-axis, the logarithm of the ranks of the corpus words; and on the y-axis, the logarithm of the frequencies and you can see that it's relatively close to a straight line. A Zipfian distribution says there is a relatively straight-line relationship between these two things, meaning there's very many words that occur infrequently and very few words that occur very frequently: That's essentially what we're seeing here. The graph here is not that important, it's

more important that you know how to create a frequency list but just to show you that even this little example that we observed here kind of at least seems to follow that particular trend.

The steps are, before we look at an extension of this, just loading the file, maybe cleaning it up a little by removing unwanted annotation, and then third, splitting it up where you think something is that is not part of a word—in languages like English that might be tricky, because a lot of things may or may not be considered part of words—then, once you’re done with the splitting, you count and sort and display.

Now, the more interesting question or the more interesting parts, of course, will only be possible if you can also do this for corpora that contain multiple files. That’s what I want to show you next.

This one is a little bit more involved, and it introduces one new thing, but everything else actually stays the same. So the one new thing is that if you have a corpus that consists of multiple files then what you don’t want to do is you don’t want to load every single file manually, extract the words, and then maybe save them somewhere, and then do stuff with them later. In this case, the Brown corpus consists of fifteen files—that’s kind of the upper limit of where you want to do it manually, because that’s already going to be pretty annoying: load the first file, and then do what everything we just did, and then you do the same thing again with the second file, with the third file ... Obviously, that’s going to be super annoying once you have more than fifteen or even ten files. Now, a lot of corpora consist of thousands of files, of tens of thousands of files. So obviously, we need some way to automate this. So programming languages, obviously, that’s what they’re all about, automating things that would be too tedious and error-prone for humans to do.

So what we need to do here is we need to use what is called a *loop*. A loop is a way to make R, or any programming language actually, do something more than once, namely a user-defined number of times. In this case, the stuff that we did to the first corpus file just now, we want to do that to every file: We want to load every file, strip the line annotation from every file, split up onto words of every file and collect them and count them. All that stuff is done over and over again to every file. All this stuff that gets repeated over and over again gets put into a so-called loop, and that requires sort of two preparatory steps.

The first step is this one: If we have several corpus files, we first need to tell R where they are so that R knows ‘these are the files that I’m supposed to be dealing with in some way’, and this is what I’m doing here: I’m creating an object called `corpus.files` and `corpus.files` is the content of a directory, or a folder as you might call it on your computer, namely, what folder? The folder that is called `Brown_tagged` in this folder and `[[then I]]` retain the complete

path to the file. Let me show you what that looks like in R, so that you have an idea.

I just ran it, and now what `corpus.files` contains is this. The file `Brown 1_A`, which is in this directory, which is in this directory; the second file of the Brown corpus, `Brown 1_B`, which is in the same folder in the same folder, and so on. Now we have all fifteen files of the Brown corpus and R knows, 'ok, this is where they are'. Now what we want to do is load—we want to create something that first loads this, cleans it up, extracts the words, stores them somewhere, and then does it again here, does it again here, until you've loaded every single file, cleaned it up, extracted the words out of it, and stored them somewhere. But in what I just said, there's already the second thing that you need to consider here.

The second thing is: if you load the first file within that loop, then you clean it up, then you extract the words, and then you have to store them somewhere, because you want to add the other words from the other files to that. So basically, what you need is you need sort of a bucket, so to speak, or a collector that will, over time, sort of collect all the words from every one of the corpus files. I mean you can think about it as if you're cooking. You have a pot on the stove with hot water in it and you peel one potato, and then you throw it in, and then you peel the second potato, and then you throw it in as well and the third potato, and then you throw that—it's the same thing here. You take one corpus file, you do stuff with it so you have the words and then you put the words somewhere for later, then you take the second corpus file, prepare it, work with it, and put those words into the same thing for storage later—that's what we need to do. In order for this to work, what we need to do is, we need to create a container or a collector or a pot for all the words from all the files to store them in. So I'm calling this thing `all.corpus.words` because it will not just contain the words from one file, but from all fifteen files.

Then we will use a loop. A loop, one of the most widely-used loops in R is defined like this: We have the word `for`, then we have something in parentheses, and then we have something in curly brackets. The stuff in parentheses has something, then the word `in`, and then something else. What does that do? It makes R do something multiple times.

Let me actually start with that one. The stuff that is supposed to be done multiple times is the thing between these curly brackets. The `...3`, that's what to do this many times. Now, what of these two other things here? This thing `[...1]` is a name that you can assign freely. You can call it whatever you want. Then this here `[...2]` is a sequence of values. Most of the time, it's a sequence of values from one to however many times you want something done. So in our case, we have fifteen files, so what we still want to do, loading them, cleaning them up, and extracting the words, ... we want to do that fifteen times, namely

once to every single file. Again, this `[[...1]]` is any name you want to give it, this here `[[...2]]` specifies how often something will be done, and this `[[...3]]` is what will be done that many times. In our case, it's going to load the file, delete the line annotation, extract the words, and then put them somewhere for later.

So that's what I'm doing here. I'm calling the first thing `counter`, because that's what it does: it counts how far have you proceeded, `in`, and then the sequence is defined as `1:15`. `1:15` just means the integers from 1 to 15, it counts from 1 to 15. `1:3` would be 1, 2, 3. `1:4` would be 1, 2, 3, 4. What that means is that R reads that line, it sees the opening curly bracket here, and then what it does is the following: It defines this thing `[[...1]]` to be the first element of that sequence. When R arrives here at that curly bracket, then it makes `counter` be 1, because that's what it starts with. So when R is here, internally—it doesn't show you that, but internally—it has said `counter` is 1 now. So now look what it does: It defines an object `current.corpus.file`, which is the result of loading, and this is all the same as before.

The only difference now is this, `corpus.files`, and then in square brackets, it says `counter`. So what this square-bracket notation does is, it says, 'of this thing here, which is a list of fifteen file names, give me this one'. The thing is, on the first iteration through that loop, when R arrives here, it defines `counter` as being 1, so that means on the first iteration, `current.corpus.file` is the result of scanning, of the `corpus.files`, the first one, which is text separated by line breaks, don't give me any output. So now `current.corpus.file` contains the contents of the first corpus file. Then, we delete line annotation at the beginning, we delete everything until the first space, and then we split it up into words. I'll talk about this in a moment. `current.corpus.file` is the first corpus file's content, we're going to clean it up a bit—I'll talk about this in a moment—we're going to extract the words—I want to talk about that in a moment—but then look what happens here: `all.corpus.words` is defined to be, and now the collecting happens, now put another potato in the pot happens: `all.corpus.words` is the result of combining the old version of `all.corpus.words`, plus the words from the `current.corpus.file`. What does that mean? Look, at the beginning, `all.corpus.words` is an empty vector. That's what this does: it sets up an empty structure to put things into. So here, this is empty, and then we create a vector called `current.corpus.words`, which contains the words from the file. Then, we make `all.corpus.words` the content of what `all.corpus.words` was before, which is nothing at the beginning, plus the currently created words. When R then gets to this closing curly bracket, which is the one that closes this one up here—all of this is the stuff that should be done multiple times—then, at the end of this, `all.corpus.words` contains the contents of the first file, and now the following happens: R goes back to here to the opening curly bracket and now

it kind of does the following. To anthropomorphize here, it asks itself, 'have I been, it checks sort of, has `counter` been every one of these values?' After the first iteration, the answer is no, `counter` has been 1, but it hasn't been 2 or 3 or 4 or anything like that. So R says, 'no, I haven't, `counter` hasn't been that, so `counter` is now set to the next value in that sequence. So now `counter` is set to 2, which means that when R gets here, now it loads `scan(corpus.files ...)` [2], the second. Then it cleans it up, and then it extracts the words. Then what happens here? `all.corpus.words` now contains the contents of the first file but now we change it to be the contents of the first file plus the just created contents of the second file. So `all.corpus.words` will get bigger and bigger, on every iteration. At the beginning, it's empty, it contains nothing. After the first iteration, it contains the first corpus file. Then after the second iteration, at the end of that, the second corpus file starts, and then the third. So this thing starts being empty, and at the end, it will have one million words, that's what the Brown corpus has. Is that clear, that building up of it? In R circles, that's actually not a very elegant way to do it, but it works fine here because this is such a small corpus.

Now what does the stuff in the middle do? To explain that, we need to have a brief look at what those corpus files actually look like. I'm going to show you that here. We're now working with the tagged version of the Brown corpus. That one looks like this: It, too, has some line initial annotation we will actually not want. There's always this pipe symbol (I have no idea why), then this little identifier here, then the colon, then the line number, then the space, and then there's the actual content of the corpus. Same here. So that's one thing we need to know. Second thing you need to know: you can see every word is now followed by a part-of-speech tag. The part-of-speech tag is separated from the word with an underscore. So *grand* and then an underscore and then the ADJ tag; *jury* and then the underscore, and the NN tag. *Took*, an underscore, and then a verb-in-the-past-tense tag [VBD]. *A*, underscore, an article tag, and so on. Also note, look at this: Punctuation marks are tagged as well and, uncharacteristically for English, there's a space in front of them. Usually in English, you don't have that. Usually you write a word, comma, space, and then the next word. In this version of the Brown corpus, every unit is separated from another by a space, even if in the real text, there wasn't one. But so that means two things, and actually those make life easier for us in a way.

The first thing that it means is that the cleaning up needs to be different. As you can see here, the stuff at the beginning we want to get rid of isn't always—like in the last example, it was always nine characters; here that is not the case. Here, it's one, two, three, four, five, six, seven, eight, nine, ten. But then here it's one, two, three, four, five, six, seven, eight. We can't say 'get rid of the first nine characters', it's not going to work. But if you look at this, can you see what will

work? Start at the beginning of the line and delete everything until and including the first space, that'll get rid of all this, but here it'll stop here. The cleaning up will now not be the first *n*, nine, characters—it will be everything until the first space. That's the one thing that gets changed.

The other thing that gets changed as well: This is now a tagged corpus, which means we can identify words on the basis of their tags. If you look at this, how would you define a word to someone who doesn't speak English, doesn't read English? [Here Prof. Gries interacts with the audience] So basically, words are 'the stuff before underscores', but you need to add one thing, otherwise, you're in trouble. [...] Imagine, so we've already cleaned up the corpus which means this is gone, including the space. So that means this *in* here doesn't have a space in front of it. [interaction] but there is a more elegant solution and the more elegant solution is to say 'a word is a sequence of things before an underscore without spaces'. So it's anything that's before an underscore, but is not a space. That means, this is before this underscore, but it's a space, so this one doesn't count. So a sequence of non-space characters before an underscore, that's going to be a word. A sequence of non-space characters before an underscore, that's a word, which means this will not be. Again, we don't have the time to discuss regexes here in as much detail as they would deserve, but this is what is going on here.

Like I told you before, this is the cleaning up part. `current.corpus.file` is what you get when you replace, and now this here means 'everything till the first space'. Again, I can't discuss regular expressions here, but that's what it does. Everything till the first space, you replace by nothing, so that gets rid of this line-initial annotation. Then, once that is gone, a word is defined as 'anything that you get when you take out the 'tags. What I'm doing here is, I'm splitting on an underscore followed by stuff that may be including one space. It looks at this and see, 'oh, here's an underscore and there's stuff after it, and here I stop'. This gets removed and the *in* remains. Then it goes here and it sees, '*the*, blah, blah, I don't care, oh, here's an underscore and that stuff after it until the space, so I'm deleting this', so *the* remains. "'L", "a", "s", "t", I don't care, ah, there's an underscore, so delete it until the next space', so we have *last*.

That's how it cleans it up. Basically, a lot of times when you work with annotated corpora, if you want to get the words, it's useful to use the annotation as a way to get the words. That is particularly useful if you have carefully annotated data [...]

Look at this. So this is not the greatest annotation ever. Here, this corpus has been tagged in a not-so-great way. Most people would agree that *according to* is one word. It's a multiword unit with a space in between. If you look up how many other words occur after *according*, there's not much, you know *to* is pretty much all there is. In some versions of the British National Corpus, for



instance, *according to* gets one tag so they recognize that's a multiword unit. Same with things like *because of*, for instance, they would not tag *because* and then *of*, they would tag *because of* and so that's really nice because if you then use the tag for splitting, you actually get *because of* as a word and *according to* as a word. Here, with this annotation, we will actually get *according* as one word and *to* as a separate word, if we want to find out the frequency of *according to* we have to do that ourselves. But so in many cases, when the annotation is really nice, being able to use it to get at the words will give you advantages just like that. Same with *in spite of*, that could be tagged as one word. I mean what else is going to happen after *in spite*, I mean it's going to be *of*.

Getting back to this then. Once we're done with this, we have a vector called `all.corpus.words`. We'll be here and we have that vector. Let me just show you in R for those of you who are not running this here, how quickly this is actually executed.

What I'm going to run now does just that: It loads every file of the Brown Corpus, it strips out the line initial annotation, splits everything up after tags, I mean using the tags, and collects the words. Again, it's a small corpus, one million words, but this is how fast it is. Done! So every file was loaded, and all these operations were done. Now we have this thing called `all.corpus.words`, which is a large character vector. Now we can actually see how many words the Brown Corpus has. If we proceed like this, about 1.1 million words, according to this annotation. You can see it really didn't take any time at all.

Then, someone said we don't want to count the punctuation marks and things like that: We don't want to know how many commas there are and stuff like that. So the next thing we might do is, we are using this regular expression—again, that means letters—so we say, `all.corpus.words`, that still contains punctuation marks and all these other characters that we don't want, so now we say `all.corpus.words` should be the result of finding letters in `all.corpus.words` and values `TRUE` returning them. What this does, it looks at all these 1.1 million character strings, and it retains only those that have at least one letter in them. So if there was *Ph.D.*, if that was tagged as a word, then this would retain it because it has a *p* in it and an *h* and a *d*, but anything that's just a comma, just an exclamation mark, just a colon, will be thrown out.

Then after that we do the same thing as before. We create a table of the words and sort it, and then we look at the top thirty, and maybe the bottom thirty.

Again, it kind of looks like what one might expect: 1.1 million words and again, *the* is approximately six to seven percent of all the tokens in the corpus, and followed by *of* and *and*, *in*, a little bit further down this time. But still we see all the function words that we would expect at the top of a frequency list. We see all sorts of weird stuff, probably proper names or something like that as well here at the bottom that are really, really rare.

Again, we can see that if we plot this, now that we've looked at more data, the line is actually even better behaved, so we do have a relatively straight line that confirms that the words in this corpus are Zipfian distributed: very many rare words, very few super frequent words.

Like I said, pretty quick run through. No one would expect you to learn a programming language from just a few hours, but I hope the general logic has become clear. Basically, you always need to know your corpora well: What is the format they come in? What is the format of their annotation? Is there a header that you may want to delete? Are there certain files you don't want to consider and things like that. But once you've done that, the overall structure is very, very often going to be something like that. Nearly all corpora come in multiple files so you always will have some sort of loop that loads every file, then processes it in some way that is useful for what you want to do later—in this case, clean up and isolate the words. Pretty much always, there needs to be something like this, namely, when you're done with processing, this file stored the results somewhere for later. So over time, you will pile all the results from every file into this container, into this collector structure, and then that's what you can work with.

Again, to remind you, this looks unduly complicated, but you need to bear two things in mind. So first, like I said with the characters, looking at how to split up words and stuff like that, that obviously is something you need to do with any corpus and with any software you're using. Don't fall into the trap of believing that WordSmith or AntConc or Monoconc are so much easier, you don't have to worry about these things there. Yes, you do. You're just letting the program worry about that and hoping for the best, which might obviously not be particularly thorough.

Second, a lot of corpora come in formats that these normal tools can't handle very well. So, one advantage of learning a programming language to do this is that you can work with any corpus. I mean WordSmith and AntConc and stuff, they work well on a small set of corpora that have a certain kind of format. But they don't work very well on anything else. Anything that's just a little bit out of the ordinary, you can actually not process very well with these ready-made tools. With a programming language, again, whatever corpus you're looking at, and even if it's one you compile yourself, you'll be able to work with it in a much more advanced way than you will with such ready-made tools. I mean especially in a language, like in Chinese writing, you don't have spaces between words: Obviously, you'll need some way to split up consecutive text into tokens. That, too, is something that a ready-made tool will probably fail at on many, many occasions.

Alright, why don't we leave it at that? Let's see whether there's any questions.

## On Recency and Dispersion

### A very brief recap

- I mentioned earlier this week that (token) freqs alone are maybe not as important as much work in psycholinguistics & cognitive/usage-based linguistics has assumed – we saw
    - doubts from Schmid regarding the nature between freq & entrenchment & this quote "frequency is one major determinant of the ease and speed of lexical access and retrieval, **alongside recency of mention in discourse**."
    - results from Baayen and others that seem to indicate that frequency-as-repetition is not that important
    - quotes from Ellis et al. pointing towards other factors
- Practice promotes proficiency (eg, Anderson, 2009; Bartlett, [1932] 1967; Ebbinghaus, 1885). **Learning, memory and perception are all affected by frequency, recency, and context of usage: The more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it, and the more fluently it is accessed** (Ellis, Römer, & O'Donnell 2016:45f.)

FIGURE 1

To start with a brief recap, again, just to make sure that everyone's on the same page: So what I talked about before, yesterday, is that token frequencies alone in particular but maybe frequency, even in general, is not necessarily the important predictor or the important driving force among many things that a lot of work in psycholinguistics, but also in cognitive and usage-based linguistics, has argued.



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.g6u360>

We've seen cases, where, for instance, even proponents of the role of frequency and the role of entrenchment have argued that there are some doubts about how far we can push this notion. Remember the quote that I provided from one of his (Schmid) articles, "frequency is one major determinant of the ease and speed of lexical access and retrieval"—ok, but then qualified by saying later—"alongside recency of mention in discourse". It is worth pointing out—that's not to be meant as a critique—but it's worth pointing out that sort of in that paper, he does say that "alongside recency of mention in discourse", but then actually doesn't discuss that. So in a sense, in that paper, it seems he realized that this is important, but it doesn't receive any further discussion in that paper and this is something we want to address today.

Secondly, I talked a little bit about the result in one paper by Harald Baayen where he showed that frequency, once a lot of other things are controlled for, doesn't do much, and I mentioned a few other variables that, at least when you compare them to frequency, seem to have a much greater degree of predictive power. So in a nutshell, it seems frequency as a repetition-counter, the way it is often talked about in many cognitive or usage-based approaches, is actually not that important. I then also mentioned briefly at one point this quote from Ellis, Ute Römer, and Matt O'Donnell (2016:45f.), where the point is that they point to a variety of other factors, the first of which we want to discuss today. So again, remember they said "learning, memory and perception are all affected by frequency"—ok and then—"recency and context of usage". Since today's topic is recency, "the more recently we have experienced something, the stronger our memory for it, and the more [fluently it is accessed. That's basically today's topic, the different manifestations that recency can have and the different ways in which it can impact linguistic choices made by speakers as they produce language online.

Now if you look at recency, there are essentially two different ways in which you can take this psychological or cognitive concept and measure it or operationalize it corpus-linguistically in two ways. One would be the short-term perspective on recency, which might be discussed under notions such as priming or persistence, or if you want to come at it from a statistical angle, autocorrelation, the fact that the presence or absence of something is correlated with itself, its prior presence or absence of that very same thing. We will actually see in a moment that the situation is unfortunately much more complex and dangerous than just that.

The second way to look at recency is sort of in the long-term, namely using the notion of dispersion. I already showed you a little bit what dispersion is about, namely this idea that, for instance, words or constructions, any unit you want to look at in a corpus, can be distributed very evenly within that corpus

## Today, we will talk about the 2nd crucial mechanism in this quote: recency

- Recency can be seen as being manifested corpus-linguistically in two ways
  - short-term: priming/autocorrelation
  - long-term: dispersion
    - across speakers (ie often files) (recall Dąbrowska 2016)
    - across registers/genres/other corpus parts
- recency is hardly ever utilized outside of the context of priming in both cogn & corpus linguistics
- this is unfortunate, because we know that
  - priming/autocorrelation has a lot of predictive power
  - aggregated freqs disregarding dispersion mean little
- let's unpack those things ...

FIGURE 2

or very unevenly within that corpus. So that obviously covers a much longer time span than something like priming. I mean priming is supposed to be long lasting, but even long-lasting accounts usually go back maybe like ten, fifteen minutes or something, a bunch of sentences, but usually no more. Longer term dispersion of an item throughout the corpus might actually, theoretically, if it's a diachronic corpus, cover a much, much longer time period.

The important thing is that you measure dispersion across the right kind of units and I will talk a little bit about that. One particularly salient unit, of course, would be that of an individual speaker, which, if you translate that into corpus linguistics, often corresponds to a file. You know a lot of times the data from one speaker are in one file, the data from the next speaker are in the next file, and so on. So obviously, this is one way—including dispersion in your kind of analysis would be one way to do justice to this idea that there will probably be a lot of individual variation along the lines of what Ewa was saying in her overview paper (Dąbrowska 2016), and corpus linguistics with dispersion can take that into consideration. Other ways in which dispersion might be measured might be across registers or genres, or other corpus parts that are defined a lot of times by corpus compilers. It might be interesting to see for some application, not so much for cognitive ones, how items are distributed across different registers compared to maybe one central overall register or something like that.

Now, a big problem is that outside of priming, both cognitive and corpus linguistics completely underutilize or under-consider this notion of recency. You will find some cognitive or usage-based studies that look at that, that included, for instance, studies of alternation phenomena of the type that I will talk about later. But I usually don't see a lot of mentions of this, although everyone, I guess, cannot really help but agree that something like recency, of course, is relevant to processes of learning, acquisition, and online processing. So that's something we will want to address. It's unfortunate because as you will see later, priming in the sense of autocorrelation has a very high degree of predictive power already. If you want to, for instance, look at the degree to which constructional choices are predictable on the basis of linguistic and other kinds of contextual predictors, I will show you one example at least, to anticipate that already, where you can make the correct prediction about what a speaker will do just by looking at what they did last time and you can get classification or prediction accuracy in excess of eighty percent. So any linguistic predictors that we so often talk about—like animacy, length, givenness, all these things that determine syntactic or structural choices—I'm not saying they're all useless, because obviously they're not, but compared to priming, they really sometimes seem to take a subordinate role, because statistically speaking, at least, this [[priming]] is already extremely strong. So remember the example that I talked about the car's mpg example, right? If you postulate that a certain predictor like animacy or length or givenness or definiteness or anything like that plays a role for the choice or not-choice of a linguistic item, then you can only really say that that predictor is strong if it does better than everything else we already know. Well, if everything else we already know includes something like priming, and that can already get eighty percent right without anything else, then, of course, that raises the bar quite a bit in order for us to see whether our linguistic or contextual predictors are actually still doing something.

Second, a lot of the work actually in linguistics in general that uses corpus data uses them in what I call here an aggregated way. So we talk about the frequency of a word in the British National Corpus or the frequency of the word in COCA (Corpus of Contemporary American English) or something like that. But of course, what these frequencies do is they are frequencies that aggregate basically the whole corpus in one number, what is the overall total for that corpus, when in fact, as we know, if there's a high degree of speaker variability, then you know any average we cite basically on the basis of corpus as a whole will be coming with such a huge degree of variability that actually you're probably not really saying very much unless you filter that variability out. So what I want to do today basically is unpack these things and show how these things are, on the one hand, threats to analyses that don't take these things into consideration, but on the other hand how can we maybe address these kinds of

concerns for at least some of the kinds of studies that are relevant in corpus-based approaches to cognitive linguistics.

Here's a very nice quote that can be used actually for a lot of different things in cognitive or usage-based linguistics or exemplar linguistics or whatever you want to call it, but it is fitting in the context of recency as well:

Each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other.

HALLIDAY 1991/2005:67

So the notion or the idea would be whenever you encounter something, it has an effect on your linguistic system because you process it, and for at least some briefer period of time, or even for longer, your linguistic system is adjusted because of that little perception of that little processing instance. So that is one of the reasons why it is so important that we control for recency, something that you might be interested in, might be influenced, not by everything happening at the same time, but something that happened five minutes ago. That's something that I will come back to on the basis of a spreadsheet example in a second.

### Recency as priming: what's that?

- The first manifestation of recency is priming, ie the fact that an occurrence of x increases the probability of x recurring beyond its (frequency-based) baseline
  - if you've just described a transitive scenario w/ a passive sentence, you're more likely to describe the next transitive scenario also w/ a passive than if you'd just described a transitive scenario w/ an active sentence
  - if you've just described a transfer scenario w/ a prepositional dative, you're more likely to describe the next transfer scenario also w/ a prepositional dative than if you'd just described a transitive scenario w/ a ditransitive
- words can prime themselves like that, too, and they can prime semantically related words, etc
- ie there's different kinds of priming: syntactic, lexical, semantic, phonological, non-linguistic, ...

FIGURE 3

So, recency as priming: what is that? Well, priming can be defined as follows: it's the fact that the occurrence of something, of *x*, increases the probability of *x* recurring again beyond its frequency-based baseline. Again, the occurrence of something has an effect on the probability that you will use that thing, that same thing, again in a little while. And it has an effect that usually at least is facilitatory so it becomes *more* likely to use something else again. So here's one of the classic examples that has been discussed to death in the relevant psycholinguistic literature, namely the voice alternation, so your uses of active versus passive sentences. So if you've just described a transitive scenario, where transitive scenario refers to an event where some agent acts on a patient, having an effect on the nature or the structure of the patient—"He broke the window" would be a typical example—so if you've just described such a transitive scenario with a passive sentence, you're more likely to describe the next transitive scenario also with a passive, more likely than if you'd described the first transitive scenario with an active sentence. So in general, in English, actives totally outnumber passives by a ratio of like nine to one or eight to one, or something like that in some corpora at least. But if you've just uses a passive, then the chance that you're using a passive again will not be just eleven or twelve percent, it would be way higher than that because of the recent use of passive of your linguistic system. Same thing, for instance, in the case of a dative alternation: If you've just described a transfer scenario, so a scenario where an agent gives a patient to a recipient, if you've just described such a scenario with a prepositional dative, you're more likely to describe the next transfer scenario also with a prepositional dative than if you've just described a transfer scenario with a ditransitive. So typically, transfer is described with ditransitives but if you've just used the prepositional dative, you're more likely to do that again. Or if you've just heard someone else use a prepositional dative, you're more likely to do that yourself again.

These kinds of effects we find on multiple different levels of linguistic analysis, pretty much anything that is arguably represented mentally. So we find that on the level of syntactic structures of the type that I've just shown, but we also find that on the level of words. If you use a certain word very, very rarely, but then you're in a situation where you use it once, then it's not unlikely that you will use it again relatively quickly as well and, of course, words can also prime semantically related words. The example that has always been used in the relevant literature is that of *doctor* and *nurse*. If a subject reads *doctor* on a screen, they will be faster to recognize the word *nurse* if it shows up later because of the semantic connection between the two items. So that means there's different kinds of priming: syntactic, lexical, semantic, phonological, and actually



## Recency as priming in corpora

### • How would you even recognize it in corpus data?

1	CORPUS	FILE	LINE	PRECEDING	MATCH	MATCH/LEMMA	SUBSEQUENT	COMPLEMENTIZER	LengthMatrixSubj	ComplementSubjLength
2	ICE-GB	S1A-001	#12:1B	I think	think	think	the m <> the main perceptio	absent	1	79
3	ICE-GB	S1A-001	#12:1B	'm <> um <> unbalanced <	was	be	I think the m <> the main per	present	79	65
4	ICE-GB	S1A-001	#125:1B	Uh I	was	shocked	<> I mean I wasn't shocked	present	1	12
5	ICE-GB	S1A-001	#127:1B	'powerful and moving um <	is	be	Um so one of the things that	absent	48	5
6	ICE-GB	S1A-001	#13:1B	'n something that I I saw a lot	was	be	<> that when people were <	present	28	6
7	ICE-GB	S1A-001	#2:1B	I think	think	think	the main things that I saw as	absent	1	85
8	ICE-GB	S1A-001	#28:1B	I think	think	think	that the <> what I get out of	present	1	38
9	ICE-GB	S1A-001	#29:1B	'nobody is left out of this group	is	be	Um <> the difference <> I	present	75	18
10	ICE-GB	S1A-001	#29:1B	Um <> the difference <> I	think	think	the main difference that I feel	absent	1	75
11	ICE-GB	S1A-001	#31:1B	'he work that I was involved in	was	be	Um <> and I think one of th	absent	55	1
12	ICE-GB	S1A-001	#34:1B	'top <> um physical skills <	was	be	One was that I was being giv	present	3	1
13	ICE-GB	S1A-001	#35:1B	'm all sorts of other people <	was	be	The other was that this was	present	9	4
14	ICE-GB	S1A-001	#38:1B	'e w we 're working with now	is	be	that those <> movement ski	present	77	21
15	ICE-GB	S1A-001	#48:1B	We decided	decide	decide	that we would work together	present	2	2
16	ICE-GB	S1A-001	#53:1B	I think	think	think	that would be <> that 's goin	absent	1	4
17	ICE-GB	S1A-001	#54:1B	And and I	think	think	the question can <> is <> is	absent	1	12
18	ICE-GB	S1A-001	#55:1B	I think	think	think	that the problems of working	present	1	32
19	ICE-GB	S1A-001	#64:1B	And I	think	think	that 's very much <> the part	absent	1	4
20	ICE-GB	S1A-001	#66:1B	'egin to dance with each other	think	think	Um <> so I think there is th	absent	1	5
21	ICE-GB	S1A-001	#71:1B	Some peo I	think	think	some people come initially to	absent	1	11
22	ICE-GB	S1A-001	#74:1B	'use the requirement from them	will be	be	that they dance <> that prim	present	33	4
23	ICE-GB	S1A-001	#74:1B	But I	think	think	they 'vry those people are v	absent	1	4
24	ICE-GB	S1A-001	#98:1B	'nce a week um which which	means	mean	that the the pressures on us	present	5	20
25	ICE-GB	S1A-002	#10:1C	'stressed particularly because he	was saying	say	that <> it was a lot of a lot	present	2	2

- in some parts of (cogn) ling that's well known
- eg, lg acq research carefully controls for priming
- elsewhere, not so much, which is tricky ...

FIGURE 4

also a variety of cases of non-linguistic priming, so priming that is not related to anything having to do with language at all.

Now, how would you recognize that in corpus data? Because the idea is that I talk about corpus stuff here. Well, one way to look at this would be or could be exemplified as follows.

This is a dataset I'm having that you will actually see discussed in the tenth talk. So this is a dataset where the question is whether someone used a complementizer or not. So the dependent variable is whether the complementizer is present or absent. So this is about sentences like *I thought Nick likes candy* or *I thought that Nick likes candy*. You can say either one and so if the *that* is there, then this [pointing to the column COMPLEMENTIZER] would be present, if *that* is not there, this would be absent. So the question is, when do people use the complementizer and when do they not? And so the usual way of linguistic analysis that we do would be that we do a corpus analysis. Here, these are data from the ICE-GB. So the British Component of the International Corpus of English. You can see for every instance which corpus it's from, which file it's from, which line it's from, and then you see the preceding context, the match, and then the subsequent context, and then annotation for whether the complementizer was used or not. Right? And so, there's a lot of literature on

this topic. And so people have argued, for instance, that it matters which verb you're using before in the main clause. So a complementizer use is actually relatively more unlikely after *think*, and it's a little bit more likely after *contemplated*. You're more likely to say *I think Nick likes candy* and you're more likely to say *I contemplated that Nick would like candy*. So the lemma, the verb lemma, is supposed to have an impact on whether it's present or absent. That's indicated by the arrow here.

But then people have also argued that there is an effect such that the length of the subject in the main clause and the length of the subject in the complement clause also have an effect on whether people choose to realize the complementizer or not. So now the thing is that what we usually do is if we do a corpus-based study, we look at, present, absent, and so on. Then, we run a statistical analysis that tries to predict, will this be absent or present, depending on what happens in this case? Which lemma is it? How long is this first subject? How long is the second subject, and so on.

Now, priming means that this is going to be incomplete, because what it does not consider is what happened last time. This is the same file, but this is two lines above, which in speaking, it really is just maybe two seconds or something, really not that long. I mean, depending on how long those sentences are: It could be anything between like one and maybe five seconds. And so suddenly the idea is that we're so used to analyzing cases in a line-by-line way by saying what's happening here right now and has an impact on this, when in fact, it will be a sizable impact on what happened last time. The impact will be higher, the shorter that was ago. Obviously, if it was like three years ago, it doesn't matter; if it was five seconds ago, it will matter. Right? And secondly, it will be, this impact from what happened last time to now will be stronger, the more similar the last time and the current time are.

The simplest or dumbest case of this would, of course, be if you repeat yourself. If you think the other person didn't get you the first time, and you say the exact same sentence one more time, then of course it's completely identical and you'll repeat everything, but this can be a matter of degree. So like I said, it will be moderated by the distance between the two, so which is here pretty close [pointing to the two highlighted lines in the table], just two sentences. And it will be moderated by how similar are these cases to each other: the present one and the last one that has an effect, here the similarity is pretty low, right? So it's very close together, but the verb is completely different—I mean, super high frequency verb *be* versus *shocked*, which is very much more specific and much rarer. And then the subject lengths are completely different here, and also quite different here. So here, the priming effect will be actually extremely hard to predict. It's very close, so it should be strong, but it's quite

dissimilar, so it should be weak. So, that's then why we do statistics, because no one can look at this and go like, "Oh, yeah, it's going to be like this strong"—no, you need to do the math and figure it out. But so that's how you recognize priming in corpus data. So the really treacherous thing is that what sometimes can happen, is, let's imagine for a moment ... Let's take this one here [pointing to the #48:1:B line in Figure 4]: So the complementizer was realized, it's present. And so now what can happen is that you look at this case, and actually this is a good example for that: So, usually, the complementizer is absent when the subjects are really short, especially with *I*—if you say *I think*, as a native speaker, at least you hardly ever say *that* then—so when these numbers are small, this should be absent. But now if you come to this data with a perspective that doesn't know about priming, then you're looking at this and you're stunned, like, "why the hell is this present?" It shouldn't be, this is short, this is short, *decided* is a relatively frequent word, I mean certainly not an exotic word or something like that, it's relatively short. So why the hell did the speaker put the complementizer here? The answer might be, "they did it last time", okay? *That* might be the driving force for this otherwise totally surprising choice. And that is the reason why on the first day when someone asked the question about sampling, why I said "don't sample random instances, sample per file or per speaker". Because if you sample instances, then you might sample this data point, but not the preceding one. And then you look at this and you're like, "why the hell did this happen?" If you sample all the instances from this file, then you *can* actually go back and see what happened last time. And then, well, yeah, "duh, he did it last time". It shouldn't be here because the subject is so short but maybe it was a priming effect, just a carry-over from what the speaker decided to do last time for whatever reason.

And look at this [pointing to the #38:1:B line, one before the "decide" line]. Last time, there really was very good motivation to put the complementizer *that*: super-long subject here: number of characters: 77, that's pretty long. And even here, 21, that's probably like between three and five words, so pretty long subject. So in this case, there was good motivation to put it here and so then maybe the fact that the bad motivation is here may be explained by the fact that, from last time, the small arrow says, "ok, use it again". So that's how we would see priming in corpora and that, again, I hope, makes clear first why you want to look at it, and second, why you should never sample on the level of instances, but always on the level of files or speakers.

Now, as I said, some areas and research do that look at that relatively well. Language acquisition research of course is a really important case, where you need to look into this, right? Because if a child, two years old, repeats a perfectly well-formed six-word utterance by the mother, then probably the child

## ok then, but why do we care?

- **Recency as priming**
  - priming is a threat to common regression modeling
    - as a form of autocorrelation, it amounts to a violation of the independence-of-data points assumption
    - it has high predictive power that, if overlooked, can make other predictors seem stronger than they are, ie it makes studies anticonservative
  - example: *will* vs. *going to* vs. *shall* in the Q&A corpus
    - switch % to *will* - % of *will* per speaker
      - many speakers don't alternate, but many also exhibit strong priming effects (esp. in the non-native varieties), variation of priming strengths is less high across topics than across varieties
    - predictive power: 80.9%
      - correct just by
        - choosing a speaker's last construction
        - choosing *will* as the first future

FIGURE 5

didn't put that together on their own, they're just repeating it. So you need to look at whether whatever the child says is actually just a re-babbling of what mommy just said or whether it's actually been constructed on his or her own. But in many other areas, this is really not controlled all that well, which can be problematic if only because you miss the opportunity to explain cases that otherwise seem unexplainable.

Okay, so why do we care? Well, we also care for some other reasons. One is that priming can be, if you don't control for it, a big threat to the kinds of regression analyses that are often done in corpus-based approaches. And that's because priming is autocorrelation, which violates the assumption of a lot of statistical data analyses that the data points are all independent, right? Regression models and many other kinds of models, they all assume every data point is unrelated to the others, which of course is exactly what priming shows is *not* the case. So obviously, there we have a problem. And then, like I said, it has a super high degree of predictive power, which means if you don't control for it, if you overlook what priming does, then other predictors will seem stronger. So there's a little bit of variability that priming could explain, but you don't include priming in your analysis, so that variability will be swallowed up, so to speak, by animacy, or by givenness, or discourse structure or something

ok then, but why do we care?

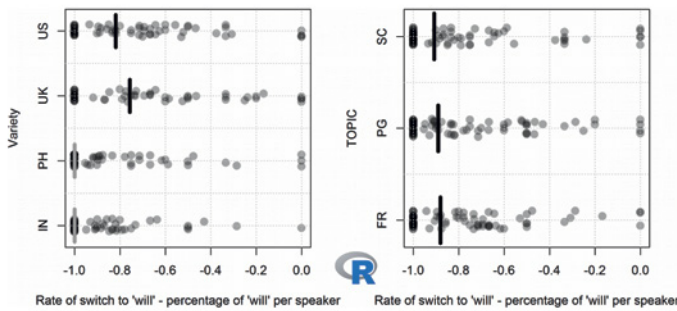


FIGURE 6

like that. And so suddenly that predictor seems way more important than it is, because actually you could explain a lot of things just by looking at priming effects, repetition effects. So what that means is studies that do not control for priming are anti-conservative in the technical sense of the term: They make things appear bigger and more important than they actually are. So let me show you an example for this high degree of predictive power.

So here's an example of an alternation, a three way alternation. So if you want to talk about future events in English, you have multiple different options. One is *will*, one is *going to*, one is *shall*, and there are others, of course. So what I looked at here in a corpus is sort of the likelihood that speakers use *will* again versus the likelihood that they switch from something else to *will*. And that is then shown here. So on the x-axis, we have their rate of switch to *will* minus the percentage of *will* per speaker, so including speakers' specific variability here. Then on the y-axis, we have two resolutions on the data. So this data is organized: It covers four different varieties of English—Indian, Philippine, UK and US—and it covers three different topic areas. So every one of these little circles here is a speaker. So first, you can see the *huge* amount of speaker variability. This is the overall average [pointing to the dark vertical thin line]. But I mean, some people are completely this extreme, some people are completely that extreme, Every range of values is attested. So this is why it's stupid to just report that one mean value, like you're blanking out the fact that

## ok then, but why do we care?

- **Recency as priming**
  - priming is a threat to common regression modeling
    - as a form of autocorrelation, it amounts to a violation of the independence-of-data points assumption
    - it has high predictive power that, if overlooked, can make other predictors seem stronger than they are, ie it makes studies anticonservative
  - example: *will* vs. *going to* vs. *shall* in the Q&A corpus
    - switch % to *will* - % of *will* per speaker
      - many speakers don't alternate, but many also exhibit strong priming effects (esp. in the non-native varieties), variation of priming strengths is less high across topics than across varieties
    - predictive power: 80.9%
      - correct just by
        - choosing a speaker's last construction
        - choosing *will* as the first future

FIGURE 7

everyone else is all over the place. But then second, you can see a lot of speakers don't alternate at all. So they always go with the same one. They always use *will*, for instance.

The predictive power that you can get if you want to predict someone's future choice, you can get more than eighty percent right by just looking at what they did last time. Of course, you have to start somewhere for the first one. For the first one, you can't look at what they did last time so there you just take *will* because it's the most frequent one. So sometimes you'll be wrong: sometimes the speaker [[*'s choice*]] will be *going to* but if you then change your prediction to *going to* from the next time, because you just look at what is it the last time, you still come out eighty percent correct across the whole dataset. I mean that's how powerful that is: Just looking at what happened last time, you really have to think about this number for a moment how scary that actually is: we're predicting future choices here without semantics, without information structure, without pragmatics, without looking at which verb is used, nothing. All the linguistic and contextual predictors that you are usually talk about, we don't even look at—we just looked at what happened last time and get eighty percent right. So that really shows you how strong of an effect this can be.

## when we plan on including the role of autocorrelation we should ...

- ... consider (cumulative) priming/learning effects – not just in corpora, also within experiments!
  - Scheepers (2003) explores long-term priming within an experiment by splitting the data into an early and a late half, but finds no significant effect w/ that
  - Jaeger/Snider (2008): *cumulativity*, "the number of primes of each structure previously encountered or produced [...]" (excluding the most recent prime)"
    - study voice & *that*-relativizer omissions in corpus data
    - find significant effects of cumulativity
  - STG/Wulff (2009): *to* vs. *ing* complementation in L2 Eng
    - use a sentence-completion experiment w/ German learners
    - find a suggestive tendency for within-subject-accumulative priming
  - Doğruöz/STG (2012): *satiation* (Francom 2009)
    - find that speakers of Turkish become more accepting of unconventional syntactic expressions over 8 stimuli
  - STG (to app): *verb-specific learning effects* in dat.alt.
- thus, we could add a predictor CUMPRIM ...

FIGURE 8

So that means, if you look at this, you have to consider a variety of additional things. So one is that you need to consider the fact that has been shown by now to be true, actually, that there is something that is called cumulative priming. That sounds very fancy, another way for that would be learning, short-term learning. As I mentioned on the first day, that actually already happens within experiments, within the twenty minutes that you subject people to maybe somewhat off distribution of stimuli. Here's one early attempted controlling for this, but honestly, I don't think they did that very well. So he (cf. Scheepers 2003) wanted to look at long-term priming within an experiment, but he did it in a very crude way at the time. I'm pretty sure Christoph (cf. Scheepers 2003) wouldn't do it like that anymore, but back then what he did is he took the whole experimental data per speaker and split it up into early versus late, [[i.e.]] the early fifty percent and the late fifty percent. And that's of course not very realistic, because we know that learning happens in an exponential curve—there's a power law of learning at the beginning, you learn very quickly, and then you level off relatively fast and the gains are low, or slow—so there's really no reason to assume that the learning takes, that the early phases is like the whole first fifty percent and the late one is the other one. Probably the distribution should be much more narrow like this (the beginning stage), and

then like that (the later stage should be much wider). And so part of the reason why he actually didn't find any priming might be because of that.

Subsequent work, corpus-based on priming, for instance, found that there is an effect of the number of primes of each structure previously encountered or produced (cf. Jaeger & Snider 2008). So they look at *that*-relativizer omissions in corpus data and they find a significant effect of cumulativeness. So the more you've used a certain structure, the more likely over the past, the more likely you are to use it again. In this study on German learners of English (cf. Gries & Wulff 2009), we found a within-subject learning effect: So we looked at *to* versus *ing* complementation, so [[like]] *he started to smoke*, or *he started smoking*, which is what that people are using. We did a sentence-completion experiment with German learners and we found a tendency for learners to more and more prefer one of the two variants, depending on how often they had used it previously in the experiment per speaker. And in this case, this is the one I mentioned very briefly before, so here we found that Turkish speakers, i.e. speakers of Turkish living in Turkey, became more accepting of the novel creations of speakers of Turkish living in the Netherlands, and over just eight stimuli, they turn from "Okay, I don't like this at all" to "Well ... why not". Eight stimuli, that's really not a lot at all. In a paper that's to appear, I found that there are verb-specific learning effects in dative alternation so if you do a sentence completion task with a German learners of English having them complete sentences either with a ditransitive or a prepositional dative, then, over time, you find that they basically regress to the mean: So verbs that are used in one construction exclusively at the beginning, they might get relaxed over time, whereas other ones are relaxed in the opposite direction. So if you're interested in or if you're aware of the fact that you should insure yourself against priming effects, you know, you might need a predictor such as cumulative priming: another good reason to study corpus data on the level of the file or the speaker; only then can you go back all the way and see what happened there.

It gets worse, though. The other thing you should consider is that priming effects can come from different sources. So there's some really cool work done by Szmrecsanyi (2005, 2006) as part of his Ph.D. dissertation back then. He distinguished what he called persistence: two kinds of priming.

One is what he called  $\alpha$ -persistence, and that's what I just talked about, basically. Namely, you use something, and because of that, you're more likely to use that exact same thing again. You use an active and so you're more likely to use an active again. Same with words: you use a word and you're more likely to use the exact same word again. The examples he discusses would be these: you have an analytic comparative, so *more something*, and then that makes it more likely to use another analytic comparative later, even though you might



## when we plan on including the role of autocorrelation, we should ...

- ... consider Szmrecsanyi's two kinds of persistence (Szmrecsanyi's 2005, 2006 word for 'priming')
  - $\alpha$ -persistence – what we called recency-as-priming
    - previous exposure to the same variable: the use of variant X will facilitate/make more likely a subsequent use of X
      - e.g., analytic comparatives prime analytic comparatives
      - e.g., *going-to* futures prime *going-to* futures
  - $\beta$ -persistence
    - previous exposure to a related/similar variable: the use of variant X will facilitate/make more likely the use of a similar/related variant of Y
      - e.g., uses of *more* outside of analytic comparatives prime analytic comparatives
      - e.g., uses of *go* as a motion verb prime *going-to* futures
- thus, adding predictors such as LASTCHOICE and LASTCHOICEWGHT is good, but not even enough since they do not consider  $\beta$ -persistence – we might need LASTSIMILAR

FIGURE 9

have a choice. You might say *more interesting* and because you just said *more interesting*, a moment later, you say *more shallow* although you could say *shallower*. [[That]] would work as well. Second example: *going-to* futures prime *going-to* futures. That's what we saw in the *will*-future versus *going to*-future example. If you just said *I'm gonna do this*, then you're more likely to say *and then I'm going to do this* as opposed to *and then I will do this*. Nothing new here.

But then he also found what he called  $\beta$ -persistence, and that is really sort of a kick in the groin for people, especially because it cuts across lexical and syntactic priming: previous exposure to a related or similar variable. So the use of variant x of something will make it more likely that you use *something similar* a little bit later. So this is really terrible because it means, among other things this: the use of the word *more* but not in an analytic comparative—other uses ... [[e.g.]] *I like this more than that*, that's not an analytic comparative, it's not *more* and an adjective, it's *I like this more than that*—that makes an analytic comparative that does use *more* more likely later. So it's not even the same structure anymore, but it's a word that is also used in a certain structure later, and then that *word*, lexical priming, makes the use of that *structure* more likely. And even worse, sort of going against what a lot of people would talk about in grammaticalization research: if you have the use of the verb *go*, it is

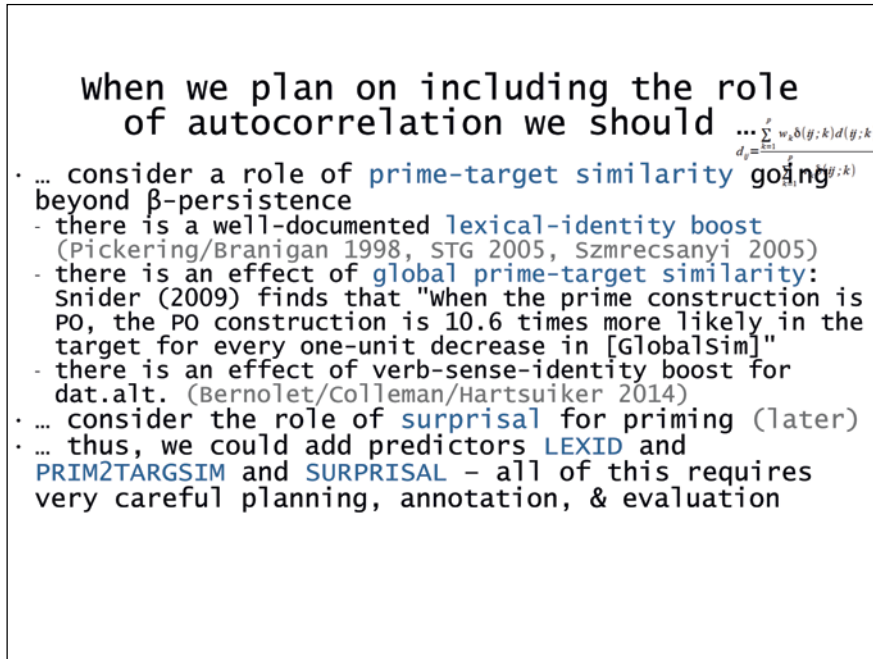


FIGURE 10

correlated with a higher incidence of *going-to* futures, even though *going to* futures, of course, don't have the literal *go* in there anymore. But he found in his corpus data, he found significant effects to that effect. So, we might need to look at what happened last time and how far away was that but we might also need how similar are these two events: Is it actually the same structure? Is it a different structure but words that are related in there? So a lot of the data that seem inexplicable at first will be explainable on the basis of this kind of stuff.

Then, like I said, you need to consider how similar a prime, the first use of a structure, and the target, the next use of the structure, are. So for instance, if you look at experimental research, there is an extremely well-documented effect that is called lexical-identity boost (Pickering & Branigan 1998, Gries 2005, Szmrecsanyi 2005), which means that if you use the same, if you have a prime that uses *give*, let's say in the ditransitive versus prepositional data, and the next stimulus is also *give*, then people are more likely to use the same construction as here than if it's a different verb. So if the two verbs are the same in the two construction choices, you're more likely to make the same choice. So that's been known like forever, but then there's also a global effect of prime target similarity. So if the prime, the first use, is similar in many other ways to the target as well, and that's what I already showed you on that spreadsheet, then

there is a huge effect for prime and target to be the same. So the more similar what happened last time to what happened now, the more likely you will find that actually also the syntactic choice in question will be the same. And the same is actually true if you look at verb sense identity. So a lot of times verbs have a lot of different senses. The priming effect is even stronger if the verb also denotes the same sense. So it's not just the same item, the same formal lemma, but also the same lemma-sense combination, so to speak.

Then we might need to consider the role of surprisal for later. That's something I want to talk about, I think, on Day 4. So all of these things, if you want to do a really comprehensive analysis of an alternation, actually you would need to make sure that you cover all of these areas to make sure that you, well, first understand the alternation best, but, second, also safe against accusations like "you forgot this", "you forgot that", all these things would need to be included. Again, I'm hoping that implicitly at least this drives home this idea that you cannot do that without statistical analysis: There's no way you look at a spreadsheet that has twenty different predictors annotated for five thousand cases, and then you eyeball it a little bit and theorize about what's in the data. That's not going to happen. At some point, I like theory as much as the next guy, but at some point, if you want to test them, you know, you do have to get down in the dirty stuff and see, is there actually an effect of this? You cannot do that by just smiling at a spreadsheet.

## Recency as dispersion: what's that?

- The second manifestation of recency is dispersion, ie the fact that occurrences of *x* are usually not evenly distributed across the parts of a corpus
- this affects
  - freqs of occurrence: *HIV*, *keeper*, & *lively* are equally freq in the BNC (16 pmw) but differ
    - re range: 62, 97, and 97, of 100 equally-sized corpus parts
    - re Juilland's *D*: 0.56, 0.87, 0.92 respectively
  - freqs of co-occurrence: verbs most attracted to the imperative in the ICE-GB: *see*, *let*, *look*, *fold*, *worry*, *listen*, *take*, *remember*, 5 more, *process* (15), but *fold* & *process* in imperatives occur in only  $\frac{1}{500}$  files ( $D=0$ )
  - everything: dispersion affects every single kind of frequency you can get from a corpus

FIGURE 11

All right, so much for priming. What's the other kind of auto-correlation? And here, I think there's also a typo that you might want to correct. So for those of you who have the book, I think it says *priming* here for you, just change that to *recency*.

So, like I said, the second manifestation of recency is dispersion, so the long-term one, the fact that the occurrences of something are usually not evenly distributed across the parts of a corpus, and this has a lot of effects. It affects, for instance, frequency of occurrence. So this is a classic example that has been repeated over and over again. I think I actually mentioned it when I was here five years ago because it is so well known: So these three words, *HIV*, *keeper*, and *lively*, they're equally frequent in the British National Corpus with a frequency of approximately sixteen times per million words but they differ a lot in terms of that dispersion. The simplest way to measure dispersion, not that I'm recommending that, but this is what was reported for this, would be the range. So you divide the corpus up, for instance, into a certain number of parts, and then you count in how many of these parts is the word attested. So *HIV* is in 62 out of one hundred equally sized parts. But the other two, although they have the same frequency, are attested in nearly all of those parts. So this word *[[HIV]]* is more specialized, it shows up in a way *[[that occupies]]* smaller number of corpus parts. If you compute a measure of dispersion that a lot of people think is good, and I will show you later that it's not, then Juilland's *D* is the same way: for *keeper* and *lively*, it's way higher *[[their Juilland's D is way higher]]*: these two words are more evenly distributed than this word *[pointing to HIV]*, which is more specialized. So any frequency of occurrence will be affected, or could be undermined, by dispersion.

The bad news continues: That's actually also true of frequencies of co-occurrence, so something happening given something else has already happened. Here's an example: If you look at the imperative construction in the ICE-GB and you measure how strongly words are attracted to that construction in a way that I will discuss in Talk 6, I think, then you get this ranking of words: *see*, *let*, *look*, *fold*, *worry*, *listen*, *take*, *remember*, and so on, some other words, *process* and stuff like that. So if you look at that, that makes a lot of sense: The fact that *see* or *let* ...—I mean *let* in particular as an imperative verb—would make a lot of sense, because that's all sorts of cases, like *let us do*, and something like that. Most of those *[pointing to worry]* will probably be *Don't worry*. Most of these *[pointing to remember]* will probably be *Remember to do something*. So if you look at this list of words, it's like, "Yeah, totally obvious", but there's two freaks: One is *fold*, and one is *process*. Why the hell would *fold* or *process* be so strongly associated with verbs—I mean, why would *fold* be more strongly attracted to the imperative than *worry* when obviously *Don't worry* is nearly a single expression. Well, for these two, they occur only in one out

of five hundred files in the imperative, i.e. they are super-specialized: the *fold* one, one of these five hundred files, is from a book on origami, that's that, that explains why that word is so super frequent there. The *process* here—and this shows you how important senses are—this is actually not like data processing or something, that's from a cookbook: *process the eggs*, you know, and then do scrambled eggs or something. I don't cook, I don't know. So super specialized. So you don't actually want to report these words as being super-strongly attracted because they're not representative of what happens in the corpus as a whole. They happen in only 0.2% of the corpus.

So the bad news is everything that you report on the basis of corpus data is potentially affected by dispersion, and pretty much always in an overestimated way: Frequency might seem impressive, but actually is not because the frequency is high, but only in one super small part of the corpus. So frequencies also will be anti-conservative: They will make you believe that something is relevant, when in fact it is only characteristic of a very, very small section of your actual data, which is why you need to control for it.

Now, like I said, what is this computed on? Ideally, a lot of times, you will find applications where a corpus has just been split up into a certain number of parts, like the study on the previous slide here [going back to Figure 11]. You

### Recency as dispersion: what's that computed on (ideally)?

- Note: most (of the too few) applications of dispersion measures are based on dividing the corpus into parts that are linguistically irrelevant
- but corpora usually come with a linguistically meaningful substructure, which provides levels of resolution over which to compute dispersion
  - files (if those correspond to speakers/texts/otherwise meaningful sampling units)
  - registers, subregisters, genres, modes, ...
- if you do not consider dispersion, any statement about 'what's in a corpus' is
  - a generalization over parts of a corpus that may be valid, but also ...
  - a generalization over parts of a corpus that
    - hopes that the H0 of equal distributions is right
    - may be terribly wrong or oversimplified if said H0 is wrong
- what, you don't believe me?

FIGURE 12

know this one, it just took the whole British National Corpus and divided it into one hundred equally-sized parts. That's actually probably not that smart, because it might even split up the corpus, I mean *corpus files*, in the middle of it: The first-million-word goes till here. Then, that's in the middle of the file, but we need to stop here, because this has to be one million words, and then it's the next one, so that's obviously not a good idea. You want to do dispersion calculations on the basis of things that are linguistically or cognitively meaningful. So ideally, what you would do is, you would use the structure of the corpus, the structures that the corpus compilers adopted or used when they compiled the corpus, to compute dispersion.

So that could mean that, for instance, you use files especially if those correspond to speakers or texts as the sampling unit, but it might also be larger things like registers, sub-registers or genres or modes, or something like that. Anything, but just a bunch of words arbitrarily picked based on a number. And if you then don't consider dispersion, any statement that you make about what's in corpus—"This is this frequent in a corpus"—is actually just a generalization over parts of the corpus that may be valid—I mean, you may be lucky—but it might also be a generalization that is completely wrong. If you say "this word is this frequent in this corpus, and that's why blah, blah, something happened", then you're just hoping that this word is relatively evenly distributed in the corpus, so that you can actually make a generalization based on that distribution. If you have this as your corpus and something is relatively frequent, but only in this little sliver of the corpus, would you then really want to write a whole paper that says how important that frequency is? Maybe not. So if you don't look at the distribution of things in your corpus, you're just going like, "please please please be right because otherwise I'm kind of screwed and just wrote a paper on something that's actually not evenly distributed".

So of course, half of you will not believe me, because it sounds like a lot of work, but it's still true. So let me give you one example here. So this is the structure of the British Component of the International Corpus of English. So every one of these levels is one that you could test dispersion on: the corpus has a spoken and a written part. And so I mean, it wouldn't be particularly smart to just look at two parts of the corpus, but theoretically, at least you could say, spoken vs. written, but then the spoken part has three sub-parts/registers—dialog, monolog, mix—and the written part has two sub-parts—printed and non-printed—and so you could look at, the frequency that you have found, is that evenly distributed across these five parts? Or is it actually, is everything just in monolog here or just in dialog here, right? Or, you could go with the sub-register. So dialog is private or public in this corpus, monolog is scripted or unscripted in that corpus, so you could look at the at the distribution of

How frequencies of present perfects change when you look at corpus parts ...

Mode	Register	Sub-register	File
spoken	dialog	private	f1, f2, f3, ...
		public	f10, f11, ...
	monolog	scripted	f20, f21, ...
		unscripted	f30, f31, ...
	mix	broadcast	f40, f41, ...
written	print	academic	f50, f51, ...
		creative	...
		instructional	...
		non-academic	...
	non-printed	persuasive	...
		reportage	...
		letters	...
		non-professional	...

FIGURE 13

whatever you're interested in on the basis of two modes, five registers, thirteen sub-registers, or actually five hundred files. Because every one of these things is instantiated by a ton of files. Just to be sure that you don't make claims on the basis of something that is super specialized. The origami book is probably from instructional writing, how to fold things, whatever, so the association of *fold* and the imperative is in a single file in this section and the rest actually—no, not at all. So again, do you want to make a big claim about the fact that something happens here? Probably not, not if it doesn't happen anywhere else, right?

So then one of my favorite graphs here. If you look at things like this, then this is what you can see or what you might miss. So this is a plot that shows you how frequent present perfects are—so *I have thought about something*—how frequent is this kind of construction in this corpus, the ICE-GB, in written vs. spoken data. The first thing to realize here is that this line here, the small dotted line, that's the overall average of the corpus. As you can see, actually the spoken data are very, very close to that overall average, they behave very much like that. The written data is actually much less than that. So if you just report that number for the corpus as a whole like everyone is doing, then you're making a good statement about what happens in this spoken data and a pretty bad

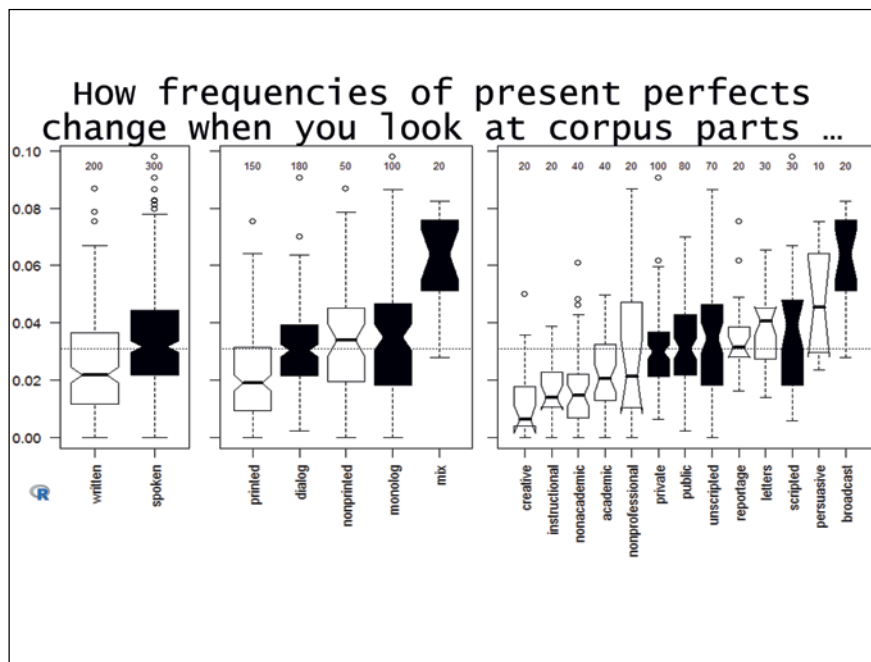


FIGURE 14

one about what happens in the written data. But it gets worse: So the black stuff here is the spoken part, the white stuff here is the written part. So this spoken average is actually spoken dialog, spoken monolog, and spoken mix. So this average, which looks like the spoken average, is actually really similar to these two, but there is a part of the spoken data that's completely different with a value nearly twice as high. But you don't see that if, like most other people, you just go with that overall average for that corpus. And then this is the thirteen sub-registers. So you can see the overall average is actually like pretty much everything in this spoken data, the four black ones other than this, which is super high. And strangely enough, the written data are actually all over the place from super small to relatively high. So really, other than "I don't want to do the work", there's really no good reason to report that one number when there's this huge amount of variability in the data, right? There's some datapoints that are even higher than this. So there's one file here that has like nearly ten percent of the verb forms are present perfects. So without looking at this, you know, these estimates become really tricky.

Here's another example having to do with a speaker-specific variability, and so what I'm showing you here in a plot is the percentages of the word *quite* spelled like this makes up in corpus data, speaker by speaker. So this is from a



## How frequencies of *quite* change when you look at speakers in corpus parts ...

- Here are some %s of the word *quite* in
  - native speaker data (EN)
  - non-native speaker data (SP, FR, NO, GE)
- here's the speaker variability ...
- most don't even use quite
- here's how sampling-dependent these results are
- many similar results: STG (2006), Callies (2013), Gablasova, Brezina, & McEnery (2017), ...

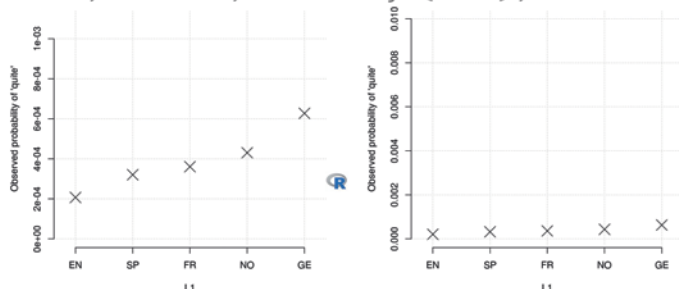


FIGURE 15

learner corpus study. We actually have native speakers, so the L1 is English, and then we have four learner groups: Spanish, French, Norwegian, and German. So this is the observed probability of *quite*. You can see that the native speakers use it least, and then the Spanish learners use it more, the French a tiny little bit more, Norwegian, also a bit more, and the Germans overuse it, it seems quite a bit. So this is what a lot of people would report [referring to the graph on the left]: You look at the corpus part as a whole, and you have that percentage. Well, let's do better than that. So first, let's adjust the scale to this: It's the same numbers, I just changed the scale here. I changed the scale, because now let's look at what every speaker is doing. And then it looks like this: Every dot here is one speaker and his or her percentage of *quite*.

So you can see that there's one German freak up here with a super high usage [referring to the extreme point of German speakers on the left-hand graph]. And then, but in general, you can see there's a huge degree of variability, but you can also see something else that's really interesting, these black things here at the bottom: Those are dozens and dozens of points on top of each other because what happens in this corpus actually is that 86.7% of the native speakers actually don't use it at all, 81% of the Spanish learners don't use it at all. So the vast majority of speakers doesn't even use the word, right? So these x-es

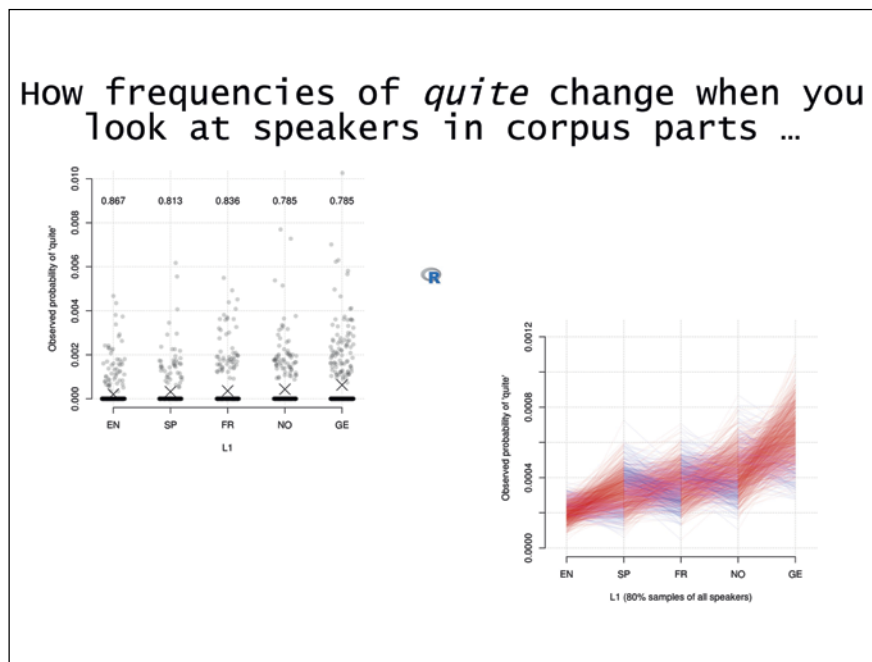


FIGURE 16

here, I mean, if you just visually look at these *x*-es [[referring to the left graph]] and do you really think this *x* summarizes this [[distribution]] very well? No, it just doesn't. Actually, you can see that these effects are extremely sensitive to random variation [[referring to the right graph]]. So what I did here is I think I generated 1000 random samples of each L1, always picking 80% of the data. So I randomly picked 80% of the English speakers, 80% of the Spanish speakers and so on and then I drew a line connecting the frequency that I found then. Then the lines are red when they go up, [[i.e.]] when the right one uses it more than the left one, and they are blue when it goes down, [[i.e.]] when the left one uses it more than the right one. And you can see, for instance, here, the difference between Spanish and French. I mean, there's a lot of red lines like this, but there's also a lot of blue lines like this. So if the sample had just been slightly different, the line would suddenly be the other way around. So again, reporting just these overall percentages without accounting for dispersion and the fact that actually more than 80% of the speakers never even use it at all would not be a good idea. This is not just me who found that, there's a whole bunch of studies out there that has looked at these kinds of things.

This conclusion I do want to show, I mean, no study that wants to be usage-based cannot at least explore different speakers' usage. Like Ewa Dąbrowska

## How frequencies can be misleading ...

- Imagine you're looking for verbs/adjectives from some frequency range in the Brown corpus (35-40 pm)
  - because you need stimuli for a psycholinguistic experiment or a vocabulary test
  - because you need words for a vocabulary list ...
- so you find these two: *enormous* & *staining* (n=37)
- but you probably didn't at all reach your goal (of finding words that are identified equally fast/accurately, that learners are equally likely to know, ...)
- *enormous*: 1 each in 35 corpus parts and 2 in 1
- *staining*: 37 in 1 corpus part
- "Language users are more likely to experience constructions that are widely or evenly distributed in time or place. When they do so, contextual dispersion indicates that a construction is broadly conventionalized, temporal dispersion shares out recency effects." - this supersedes frequency!

FIGURE 17

said, basically, you can't pretend to be sort of individually cognitive and speaker-specific, and then work with overall aggregated frequencies or percentages of a corpus.

So how can then frequencies be misleading? In a way, I already alluded to that but I want to give you one particularly clear example, even though it's an artificial one. So let's imagine you're doing something that psycholinguists are doing a lot actually: you design an experiment and you want to control for the role of frequency. You want to make sure that whatever your lexical decision task times or something are not affected by frequency, but the other thing that you think is playing a role. So you might need words in a certain frequency range, let's say 35-40 per million and you're working with the BROWN corpus. So then you might find two words that have the same frequency: *enormous* and *staining* both occur 37 times in that corpus, so the exact same frequency. And for something having to do with reading, these two words actually really great, because they also have the same number of letters. So you can say, I'm even, so "I'm holding frequency constant, and I'm holding length constant". But most of us or all of us here are non-native speakers. Do these really seem comparable? Obviously not. Otherwise, I wouldn't be using them here. Because these words are actually distributed extremely differently: The 37 instances of *enormous*

## what measure of dispersion to use ...

- Many measures of dispersion have been proposed ...
  - range: the number of corpus parts containing  $x$
  - $sd/vc$  of the frequencies of  $x$  in the corpus parts
  - Juilland's  $D = 1 - \frac{sd_{pseudo}(p)}{mean(p)} \times \frac{1}{\sqrt{n-1}}$  Carroll's  $D_2 = \frac{-\sum_{i=1}^n (\frac{p_i}{\sum p} \times \log_2 \frac{p_i}{\sum p})}{\log_2 n}$
  - Rosengren's  $S = (\sum_{i=1}^n \sqrt{s_i v_i})^2 \times \frac{1}{f}$  (with  $min S = 1/n$ )  $DP = 0.5 \times \sum_{i=1}^n \left| \frac{v_i}{f} - s_i \right|$
- Deviation of Proportions  $DP$  (Gries 2008), ie  $\text{sum}(\text{abs}(\text{OBS}-\text{EXP}))/2$ 
  - stays within its defined comparable range
  - distinguished distributions other measures can't
  - doesn't overly penalize 0s
  - has been shown to be better than the standard of Juilland's  $D$  (Biber et al. 2016, Burch et al. 2017)
- how does  $DP$  behave when applied to pseudo-randomly sampled words from the BNC sampler?

FIGURE 18

are in 36 corpus parts. So the BROWN corpus consists of 500 parts, each of which is 2000 words. And the 37 instances of *enormous* are in 36 different corpus parts. The 37 instances of *staining* are all in one of the 500 corpus parts. So no way in hell are these comparable, even though the length is the same and the frequency is the same, and it's the dispersion measure that tells you that right in your face—it's not frequency, it's the dispersion.

So, how do you measure that? As always, there's a ton of measures available. The simplest one I already told you about, range. So that's the number of corpus parts that contain a certain item. Another one would be the standard deviation, or the variation coefficient, of the frequencies of something in corpus parts. A lot of people have worked with Juilland's  $D$ , although recent work having come out in 2016 and 2017 shows that you probably shouldn't (cf. Biber et al. 2016; Burch et al. 2017), and then there's the measure that I will want to promote here is the one—surprise—that I came up with, which I call  $DP$  and it's computed like this.

So you take the observed percentages of a word in the corpus parts, subtract from it the sizes of the corpus parts, and then you take away the minuses, you sum it up and divide it by two. Sounds complex, but actually it's not. Let's look at an example here.

Example number	Exp (sizes of parts)	Obs (distribution)	abs diff	sum of abs diff	divide by 2 DP
1	0.33 0.33 0.33	0.33 0.33 0.33	0 0 0	0	0
2	0.33 0.33 0.33	1 0 0	0.67 0.33 0.33	1.33	0.665
3	0.01 0.01 0.98	0.98 0.01 0.01	0.97 0 0.97	1.94	0.97
4	0.01 0.01 0.98	0 0 1	0.01 0.01 0.02	0.04	0.02
5	0.45 0.35 0.2	1 0 0	0.55 0.35 0.2	1.1	0.55
6	0.45 0.35 0.2	0 1 0	0.45 0.65 0.2	1.3	0.65
7	0.45 0.35 0.2	0 0 1	0.45 0.35 0.8	1.6	0.8

FIGURE 19

Let's look at the first example, that's this whole thing. So let's imagine you have a corpus, I mean, this example is ridiculous, just for didactic reasons. So you have a corpus that has three parts, and they're equally large. So that means every part is 33% of the corpus. Now, let's imagine you have a word that shows up the same number of times in each of these parts [referring to the example 1]. Let's say two, two, two. Ok. Then each of the corpus parts has one third of the items. Right? And so then you compute this minus that [[and the result]] is zero, this minus that is zero, this minus that is zero. There's no minus to take away so you sum them up: it's zero; you divide by two: it remains zero. So the low values mean that a word or anything else you look at is distributed exactly as you would expect from the sizes of the corpus parts.

Now let's look at these two here, and then I'll skip the rest. So here we have a ridiculous corpus, three parts [referring to Example 3]. One is 1% of the corpus, the other one is 1% of the corpus. The third is the remaining 98% of the corpus. Unrealistic, it's just a didactic example. But now let's imagine that nearly all the examples of what you're interested in are squeezed into one of these small parts, like with *staining*: 37 items, but they're all in one of the tiny corpus parts. So then you get this minus that, and we take away the minus. This is the same. And then we get another huge discrepancy: There should be a lot of things

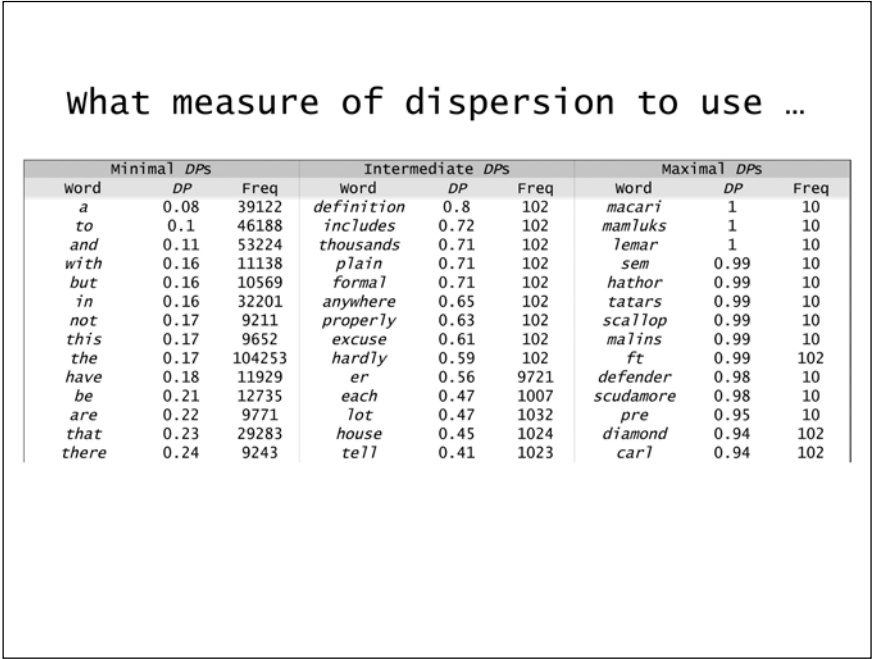


FIGURE 20

in there, but they are not, so there's a huge discrepancy. You sum them up, divide by two, and so this is very close to one, which is the theoretical maximum value. So high values mean that the word is super specialized. Whereas here [referring to Example 4] we have the same corpus, small, small, huge. But now all the instances of the word are in the big part, which is what you would expect, and so this comes out to be very small. So this DP measure ranges theoretically from 0 to 1, [[with]] 0 meaning the word is super evenly distributed [[and]] 1 meaning it's super specialized. It has some nice characteristics that you can read up on here.

Now, let's see what this measure does when you apply it to a corpus. So these are the words that are most evenly distributed in the corpus, and you can see that there is obviously a correlation with frequency: We have all the function words here that you would expect to find everywhere. So in the middle, you have words that are somewhat evenly distributed and these are all words that, for instance, as non-native speakers of English, we all know those. They're relatively evenly distributed. I mean obviously, *properly* is much more specialized than *that*, but we still know those. But now the words that are super-exotically distributed are these and several of those, I'm a relatively advanced nonnative speaker, I have no idea what that is, I don't know what that is, right? So

## what that measure of dispersion does & how it relates to frequency ...

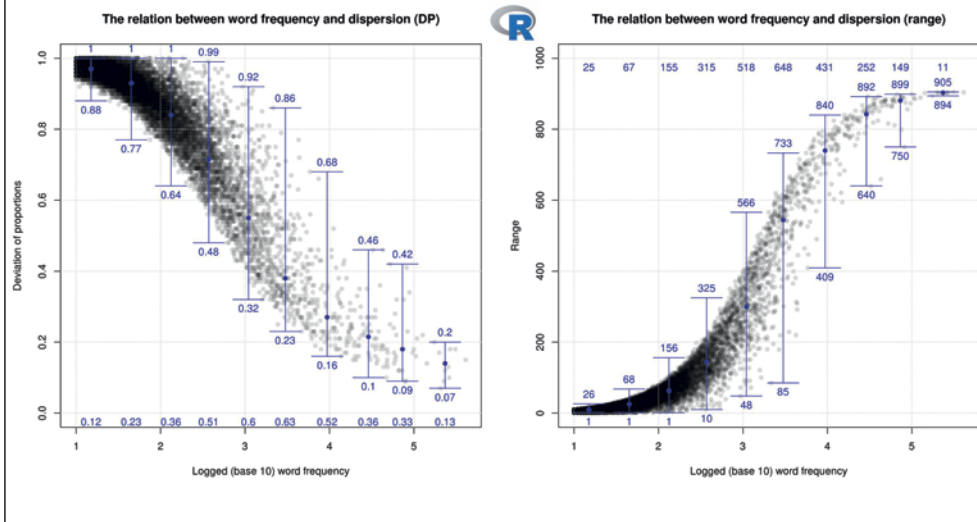


FIGURE 21

obviously, these are much more specialized and that's exactly what that measure shows so it works.

Frequency and dispersion are correlated. You saw that on Day 1, but the interesting thing is this.

So let's use this one here. So here on the  $x$ -axis, we have the frequency of a word as a log to the base of ten. So [[1 on the  $x$ -axis]] is ten, [[2 on the  $x$ -axis]] is 100, 1000 [[three on the  $x$ -axis]], 10,000 [[four on the  $x$ -axis]], 100,000 [[five on the  $x$ -axis]]. And then here we have this measure of dispersion. So, one of these here is *the*, one of these here is a mix typed word that shows up a single time or something like that. So obviously, there's a hugely negative correlation: if something doesn't happen very often, it can't be distributed very widely. But the scary thing is this: Look at the middle range where all the content words are. These are all function words. And then here is where we have frequent nouns and verbs and stuff like that, also known as 'the words that people usually run experiments with'. And that's the range of values where there's the hugest discrepancies in terms of dispersion. Words can have the exact same frequency of something between 1000 and 10,000 and be really evenly distributed nearly as much as words like *of* and *in*. And they can be super specialized, like words that you and I may not even know. Although someone might legitimately say, "but I

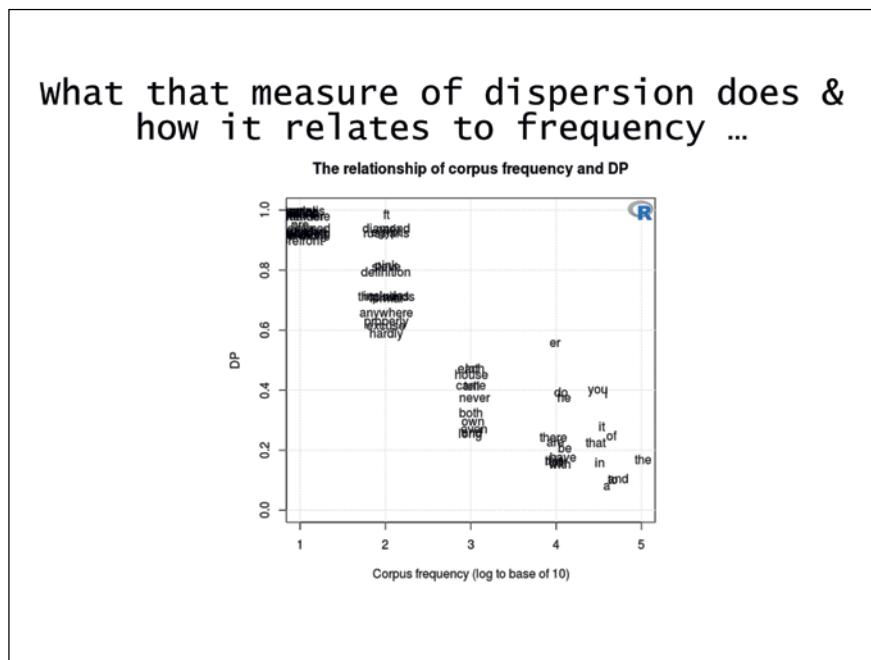


FIGURE 22

controlled for frequency”,—well, yeah, but they didn’t control for this  $[[y]]$  axis and so they actually didn’t control for much.

Here’s an example. If we look at some words: so same frequency, something as, I mean, function words like *even* and *both*, but also word like *earth*,  $[[$ which are]] much more specialized. Or here, words like *hardly* and *properly*, but also something like *diamond*, much more specific. I mean very narrowly distributed word, we don’t talk about it much.

So what we have here is a situation where we have both a huge degree of theoretical motivation to talk about dispersion, because, again, dispersion is a corpus linguistic function of recency and no one in their right mind would ever dispute that recency has an impact on processing and learning. And so here’s one very nice quote, “learning is always better when exposures or training trials are distributed over several sessions than when they’re massed into one session” (Ambridge et al. 2006:175). That sentence is the reason why students shouldn’t be cramming for an exam the night before, but evenly space it out over time so that their learning is better: better dispersion, better learning. Learning is related to separations of exposures in time and context. Here is another study that says “item’s later retrieval depends on the separations of the exposures in time and context” (Adelman et al. 2006:814). So a whole



## what theoretical motivation do we have to use dispersion?

- "Given a certain number of exposures to a stimulus, or a certain amount of training, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session. This finding is extremely robust in many domains of human cognition." (Ambridge et al. 2006:175)
- learning is related to separations of exposures in time & context (Glenberg 1976, 1979)
- the extent to which the number of repeated exposures to a particular item affects that item's later retrieval depends on the separation of the exposures in time and context" (Adelman et al. 2006:814)
- Schooler & Anderson (1997) also demonstrated that there is a power (i.e., log-log linear) function relating probability of a word occurring in the headline in the NYT on day  $n$  to how long it has been since the word previously occurred in that context. The human forgetting curve (Ebbinghaus, 1885) is rational in that it follows this trend. (Ellis, Römer, & O'Donnell 2016:37f.)

FIGURE 23

bunch of quotes—this one I'm not going to read out to you, but you have it in your handbook—that shows there's ample theoretical and psychological background literature showing that recency plays a role.

What do we have in terms of empirical motivation? Well, if you were to ask me this, my first reaction would actually be, "you mean apart from everything else I've already shown you?", but again, there's a lot of different kinds of studies. So for instance, Ellis and colleagues have shown that range has a significant predictive power when it comes to construction uptake beyond raw frequency (Ellis & Simpson-Vlach 2005; Ellis et al. 2007). The Adelman study shows that range is a better and more unique predictor of reaction times than frequencies (Adelman, Brown, & Quesada's 2006). I showed in one study that dispersion measures correlate more highly than frequencies with different times of response time latencies (Gries 2010).

So here I tested approximately twenty-five different dispersion measures and frequencies for how strongly they are related to reaction times. And so this is where frequency is, really suspiciously close to the zero line of 'no correlation'. And then there's a bunch of measures that are positively and negatively much more correlated with reaction times but as you can see, pretty much *all* dispersion measures do better than frequency.

What empirical motivation do we have to use dispersion?

- You mean apart from all of the above?!
- Ellis & Simpson-Vlach (2005) & Ellis et al. (2007) show that range) has significant predictive power above & beyond raw frequency
- Adelman, Brown, & Quesada's (2006) show that range is a better & more unique predictor of RTs
- Gries (2010) shows that some dispersion measures correlate more highly than raw frequencies with
  - response time latencies from Balota & Spieler (1998)
  - lexical decision task times from Baayen (2008)
- Baayen's (2010) comprehensive analysis mentioned earlier finds that dispersion is the second strongest of 19 predictors of lex dec times
  - yes, in that study frequency is the strongest, but
  - frequency is 91% explainable from everything else, &
  - repetition frequency does little else

FIGURE 24

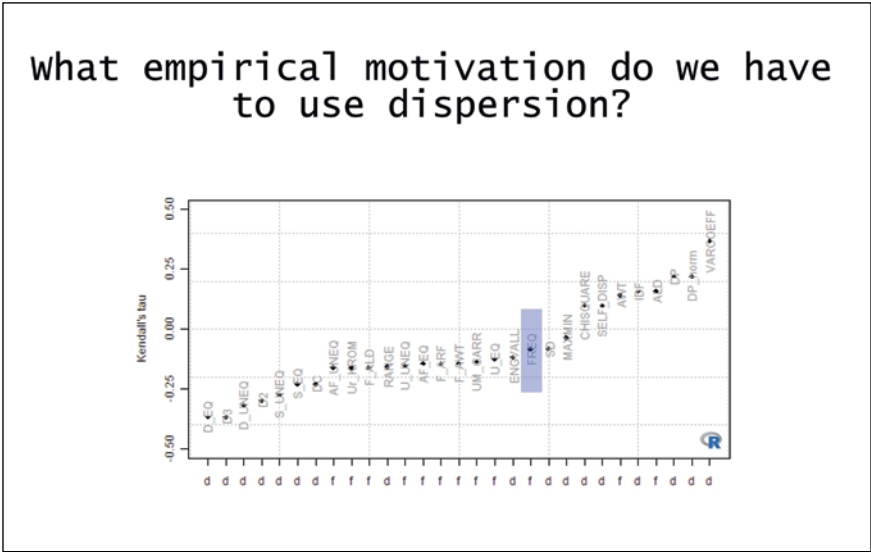


FIGURE 25

## what empirical motivation do we have to use dispersion?

- An extended example: Balota & Spieler's RT data for 2820 words measured for both older & younger subjects
  - I took 6 corpora
    - BNC Baby, BNC Sampler, BNC, BNC spoken, Brown, ICE-GB
  - computed  $DP_{\text{norm}}$  of each word type across files
  - correlated each word's RT w/ frequency &  $DP_{\text{norm}}$  to determine which predicts RTs best
- results
  - frequency is never the best predictor (despite its ubiquity, strong support for Baayen, Adelman et al.)

Deviance expl. by GAM	FREQ		$DP_{\text{norm}}$	
	young	old	young	old
BNC Baby	4.96	7.06	8.48	14.9
BNC Sampler	5.22	6.44	9.07	13
BNC	5.06	7.57	9.26	17.3
BNC spoken	4.26	5.88	8.64	14.3
Brown	4.78	6.77	7.85	13.2
ICE-GB	3.79	4.78	6.1	9.3

FIGURE 26

I mentioned Baayen's study from before: dispersion is the second strongest predictor of lexical decision times. And yes, frequency was the strongest, but that was frequency that was still correlated with everything else. If you take everything else out and use frequency as a repetition-counter, like what I've told you before, then you know 91% of explanatory power go away and it actually doesn't do anything anymore, which then leaves dispersion as the strongest predictor.

Slowly coming to an end at least. Here's another example. I did a similar study essentially. I took all words from the MRC Psycholinguistic Database and correlated and checked whether they were in any one of these six corpora, where BNC spoken is a subcorpus of the BNC. And then I computed my dispersion measure for each word and of course, I had the frequency of each word and then I checked for every one of these six corpora how well can the reaction times from people in the experiment be predicted by this corpus either by frequency or by dispersion? And the database here includes reaction times from younger speakers and from older speakers, but actually it doesn't matter. These numbers indicate the amount of explanatory power of frequency and dispersion. And you can see in every single comparison, dispersion does much better than frequency, sometimes more than twice as high. Even if you run a

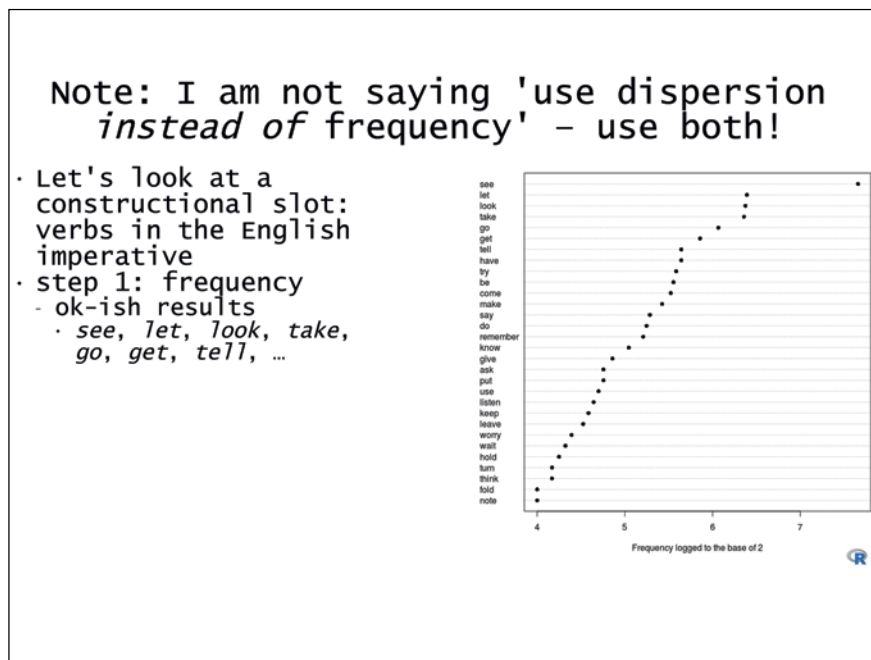


FIGURE 27

direct comparison of two things that are strongly related, this one explains what everyone is talking about—“speed and ease of lexical access” is what Schmid said—the dispersion measure explains things much, much better than frequency, so really think twice about whether just because frequency is easy whether that means it's the thing you should include.

Now, to make your life a little harder, this is the last part: I actually don't want to be saying 'use dispersion instead of frequency'. Ideally, what you would be doing is you would be using both because they are correlated, but not deterministically, right? There was a huge degree of spread in the middle where the interesting content words are. So let's look at a constructional slot, namely verbs in the English imperative, the example we looked at before: *see, let, look, fold, process, worry, remember, listen*, these kinds of things. So if you want to talk about the semantics of that construction, then you know a lot of corpus linguists and cognitive linguists, I think, would say, well, look at the frequency with which verbs show up in that construction, right?

If you do that, you get this ranking in the frequency and the frequency again is logged here. So *see* is the most frequent word there, then *let*, then *look* and *take, go, get, tell, have, try, be*, okay. Actually, it doesn't look that bad, right? Even I argue against frequency all the whole time, I have to admit that's not

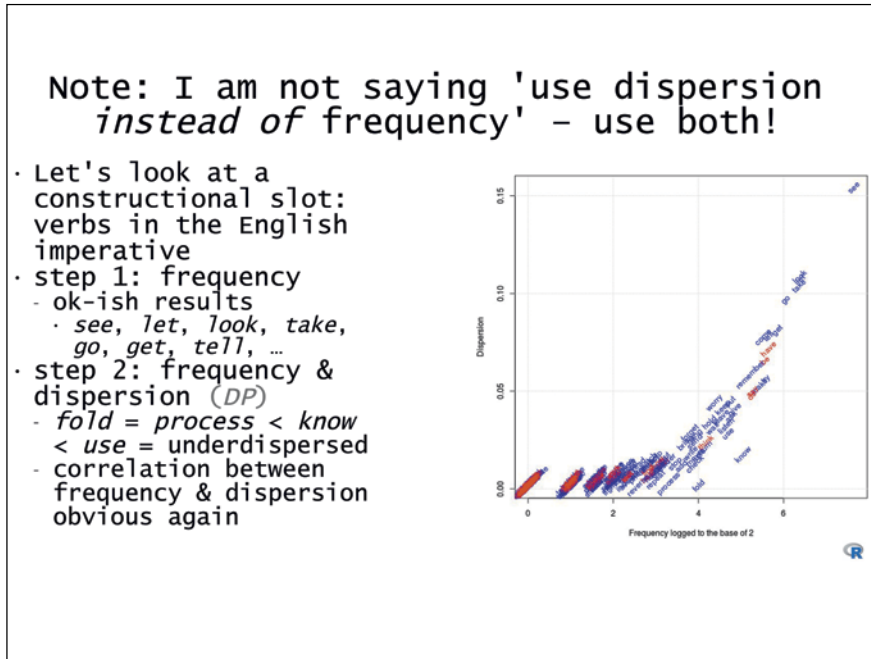


FIGURE 28

all that bad: *See, let, look, take, go, get, tell*. I can easily imagine all of those be used in imperative, although I do want to point out one thing that is maybe a little bit weird. I'll come back to that later: *have* and *be* show up here pretty highly. I mean *have* is what, like number seven or something like that or eight; *do* is also relatively highly up here. I'll come back to that, ok, file that away for future use.

But so now let's look at frequency *and* dispersion. So how frequent is a verb slot in the imperative and how often/ how well is it distributed across the corpus in that verb slot? Then we have frequency here, same scale as before and now we have this dispersion measure here. I flipped it: It was 0–1 and I made it the other way around so high values mean even dispersion, even distribution. So *see* still 'wins'. Then there's *look, let, take, go, get, come, child*. Okay, fine, but I've added one piece of information, namely the color. What does the color mean? The color means whether the word shows up in there more or less often than expected, and so *have* and *be* and *do* are in red. So that means they are actually relatively frequent in there, but they are verbs that are so frequent in general, they should have been more often in the imperative, given their overall super high frequency, they are actually less frequent in the imperative than expected. If you just look at frequency, you don't even see that, right? You

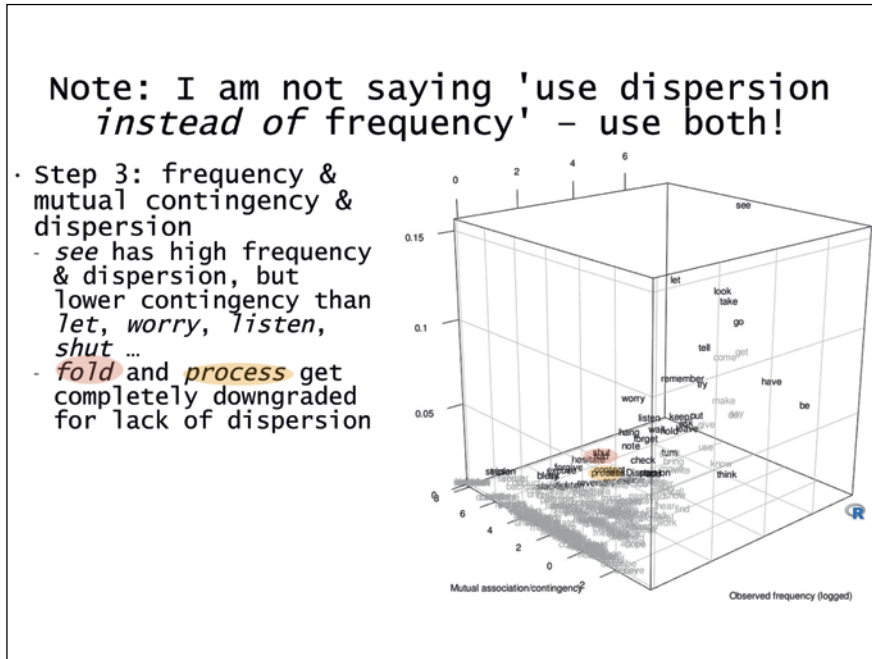


FIGURE 29

can just look at the size of that number and then be like, “oh, wow, number eight”—only when you relativize it (*be* and *have* and *do* are pretty frequent words in general), then you do realize that this is actually pretty little. But it still looks kind of fine, right? I mean we have *worry* and *remember* being much better, a bunch more dispersed than things like *forget*, *check*, things like that. But notice already what happens here? Here’s *fold* and here’s *process*, right? So relatively frequent, close to *worry*, for instance, or close to *forget*—*forget* is a good imperative, because *Don’t forget to buy the milk or something*—but already adding the dimension of dispersion makes you immediately not consider *fold* and *process*, because they’re way down here so that means they’re super specialized. So just by adding that perspective, you’ve already avoided the mistake of making these seem important. Because the words that are important, they should be higher up and to the right—these are relatively to the right or in the middle, but they are as far down as possible, namely at zero, and so you know not to use them. So just adding this little fix, makes sure that your frequency-based values are also corrected or controlled for dispersion. So *fold* and *process* and *use*, they are underdispersed. And so you wouldn’t want to use them as good examples. You wouldn’t want to use them as stimuli in your experiment for the imperative or something like that.

## Use both, but separately!

- Freq & disp are correlated ( $R^2=0.83$  BNCsp), but
- in the middle range of frequencies, words can have very similar frequencies but unequal dispersions
  - *staining* vs. *enormous* (in
  - *church* vs. *place* Brown)
  - in the 6th freq bin of BNCspkn
    - *council*: freq=4386, DP=0.72, range=292 out of 905
    - *nothing*: freq=4159, DP=0.28, range=652 out of 905
    - *try*: freq=4199, DP=0.28, range=664 out of 905
    - *whether*: freq=4490, DP=0.32, range=671 out of 905
  - the correlation between frequency & dispersion is low:  $R^2=0.08$

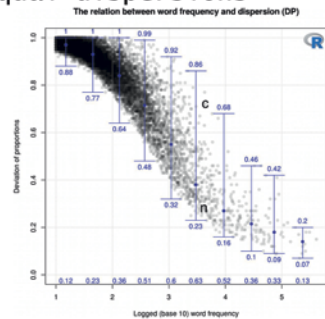


FIGURE 30

Now what if we add attraction, contingency, so the topic of actually one of the next talks, then we get a plot like this. So now it begins to be complicated, because actually we have three dimensions of information. We have frequency, we have dispersion going up, and then this side of the cube is this contingency measure that says whether something is preferred or dispreferred in there. And so you can see that the usual suspects win: *see*, *let*, *look*, *take*, *go*, you see that all the grey stuff is actually not significantly attracted. And you can see here, it's not easy to see that in the two-dimensional display of a three-dimensional cube, but you can see here at the bottom, *fold* and *process* are on the floor of that cube because they score so low on dispersion. So you do not want to use those.

So again, you saw that in the cube. You saw that in the graph before, frequency and dispersion are correlated, and actually really highly: 83% of the variability of the dispersion values can be explained on the basis of frequency. But again, that is unfortunately not an excuse to not do it because, like I said, in the middle range of frequencies, here, that's where the words with similar frequencies differ mostly in terms of their dispersion. And you can see it from some examples very well. So this is in the BROWN corpus: so here's *staining* and here's *enormous*. They have the same frequency. What was it? 37. But one of them is as evenly distributed as possible for words that are rare., that's *enormous*, and



## Use both, but separately!

- Finally, the graphs also show that forcing frequency and dispersion into a single value – an **adjusted frequency** of the kind often used in lexicography – is a bad idea because of the information loss
  - theoretically, an adjusted freq. of 35 could result from
    - freq=350 & Juilland's  $D=0.1$
    - freq= 35 & Juilland's  $D=1$
  - yes, that's a hypothetical, but
    - adj.freq. for *pull* in BNCspkn  $\approx 375$
    - adj.freq. for *chairman*  $\approx 368$
    - pull*: freq= 750,  $DP=0.5$
    - chairman*: freq=1939,  $DP=0.81$
  - in the plot on the right, all the red dots represent words with  $365 \leq \text{adj. freq} \leq 434$ , but with  $701 \leq \text{freq} \leq 1939$

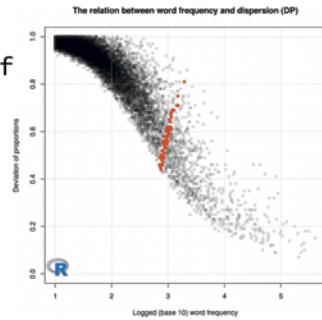


FIGURE 31

the other one is as specialized as a word with that frequency can be, and that is *staining*. I mean, look at this one here: the words *church* and *place* have nearly the same frequency but *church* is super specialized compared to something as generic as *place*. So you don't just want to go with frequency—you do have to take this into consideration as well. There's no way in hell people will react to these two words the same way in an experiment, because their frequency is the same.

Here are some other examples from the BNC spoken. So this is the word *council* and this is the word *nothing*. They both occur, for all intents and purposes, in a really similar number in ten million words. That's not a huge difference, but obviously they are super different in terms of where they show up. I mean, *council* is a relatively specialized word, it's actually amazing that it's more frequent than something as useful as *nothing*. But it's not a contradiction anymore once you realize, "well, yeah, this is what explains". You look at these two and you're like, "well, this [*nothing*] is more common", right? But the frequencies don't support that—it's the dispersions that support that. Here are some other words that are also in that same frequency range: *try* and *whether*. Again, I mean compared to *council*, *nothing*, *try* and *whether* are all like everyday words, right? But the frequencies are all the same—the dispersions do what we want them to do, namely, separate this one from the other three.



Now, final thing. So one thing you might now say, “ok, I’m going to do what some lexicography people are doing, [[i.e.]], corpus-based lexicographers. Obviously, they are corpus-based these days at least in some way, so they have a frequency of a word. They also can compute the dispersion of a word, and so then to ‘help’ people using the dictionary or linguists, they conflate the two into one value and adjust the frequency. So that’s a frequency that gets downgraded if the dispersion is too small. Seems like a good idea, but it’s not. Because first, in a theoretical example, the downgrading actually is multiplying the frequency with Juilland’s  $D$  value. So if you get an adjusted frequency of 35, then that could be a word that occurs 35 times very evenly, or it could be a word that occurs 350 times super specialized, but from that number, you don’t see which of the two scenarios it is. You lost that because you took two different numbers and conflated them into one.

And again, that’s a hypothetical example, but it’s not that far off. So, if you compute the adjusted frequencies for the words *pull* and *chairman* in the spoken BNC, they are the same again, for all intents and purposes, pretty darn close. But, look at the frequency difference between the two: Actually, *chairman* is way more frequent than *pull*, which is, again, kind of weird for spoken data, but the dispersion value clearly tells you that. This one [[*pull*]] is much more evenly distributed than that one [[*chairman*]]. So the fact that the adjusted frequencies that conflate these two data points are the same, that is not a feature, that is a bug. That is not telling you something you want to know. You want to keep those separate to see ‘this one has this frequency, because it’s really evenly distributed and this one is very frequent, but specialized’.

So in this plot here, what you can see is the same plot you’ve seen before: frequency, dispersion. So all the red dots have extremely similar adjusted frequencies. So if you conflate the two values into one, you’re essentially saying that a word like this is pretty darn similar to a word like that even though there is a slight frequency difference and a huge dispersion difference. So, don’t conflate, keep them separate because every dimension tells you something that is relevant. Remember this cognitive commitment thing? What we probably don’t do in our heads is compute an adjusted frequency. No, we have ideas about frequency and we have ideas about dispersion and recency, we don’t conflate them.

So I’m going to skip one slide and wrap up.

And again, you have that in the handbook. So the conclusions, so all the things that frequency was supposed to affect—learning, acquisition, memory processing, cognition—yes, they all are related with frequency, but recency in different versions is something that can override frequency in the short-term as priming or persistence ( $\alpha$ - or  $\beta$ -persistence), in the long-term as dispersion

## Just for kicks, note these other analogs to dispersion

- Consider the organization of warehouses or libraries: sort by frequency, but prioritize items that are currently hot, up-to-date, trendy, 'in', ...
- consider your e-mail: you talk to particular individuals in flurries of activity separated by long hiatuses
- consider the File menu of your computer software, which has 'open recent' options
- computational linguistics models: adaptive language models were introduced to account for repetition. It is well known that the second instance of a word (or ngram) is *much* more likely than the first
  - the 1st instance of a word depends strongly on freq
  - the 2nd does not: adaptation (burstiness) depends more on content than frequency

FIGURE 32

## Conclusions

- All the things that frequency was supposed to affect
  - learning & acquisition, memory, processing/cognition, ...
- are correlated w/ freq, but recency overrides freq
  - in the short term, as priming
  - in the longer term, as dispersion
- all corpus stats are at risk from such recency effects – the fact that occurrence or co-occurrence for any and all phenomena might not be evenly distributed across parts of a corpus: aggregate freqs are mostly useless for anything cognitive
- priming is highly predictive, cumulative, & moderated by distance, similarity
- dispersion explains more than freq-as-rep and should be computed over meaningful corpus parts
- but: keep dimensions of information separate
- with all that, freq effects we arrive at will be more accurate/reliable

FIGURE 33

along the lines of what you've just seen, and all corpus statistics are at risk from such recency effects, all of them. Any frequency of occurrence, and any frequency of co-occurrence can either occur really nicely distributed and be totally representative of the corpus as a whole, or it can be super specialized and you don't see it—unless you correct for this. So aggregate frequencies, where you just take everything are pretty much useless for anything cognitive. But if you look at those two things at the same time, keep them separate and don't conflate them into a single measure. Thanks.

## Dispersion: Practice with R

### what measure of dispersion to use ...

- Many measures of dispersion have been proposed ...
  - range: the number of corpus parts containing  $x$
  - $sd/vc$  of the frequencies of  $x$  in the corpus parts
  - Juilland's  $D = \frac{sd_{population}(p)}{mean(p)} \times \frac{1}{\sqrt{(n-1)}}$       carroll's  $D_2 = \frac{\sum_{i=1}^n (\frac{p_i}{\sum p} \times \log_2 \frac{p_i}{\sum p})}{\log_2 n}$
  - Rosengren's  $S = (\sum_{i=1}^n \sqrt{s_i/v_i})^2 \times \frac{1}{n}$  (with  $min S = 1/n$ )       $DP = 0.5 \times \sum_{i=1}^n \left| \frac{v_i}{f} - s_i \right|$
- Deviation of Proportions  $DP$  (Gries 2008), ie  $\text{sum}(\text{abs}(\text{OBS}-\text{EXP}))/2$ 
  - stays within its defined comparable range
  - distinguished distributions other measures can't
  - doesn't overly penalize 0s
  - has been shown to be better than the standard of Juilland's  $D$  (Biber et al. 2016, Burch et al. 2017)
- how does  $DP$  behave when applied to pseudo-randomly sampled words from the BNC sampler?

FIGURE 1

Thank you very much. I want to begin today's practice session by just recapitulating one thing about yesterday and just to make sure that we all understand how the practice session today will proceed. So, remember that yesterday I told you about different dispersion measures that exist. I mentioned Juilland's  $D$ , the fact that it's a kind of a standard, but that recent studies have shown that maybe other ones are better, and one of those that have been shown to be better is the one that I (Gries 2008) developed a few years ago called  $DP$ , and I just wanted to reiterate what that formula does, because in order for us to



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/m9.figshare.9611414>

Example number	Exp (sizes of parts)	Obs (distribution)	abs diff	sum of abs diff	divide by 2 DP
1	0.33	0.33	0	0	0
	0.33	0.33	0		
	0.33	0.33	0		
2	0.33	1	0.67	1.33	0.665
	0.33	0	0.33		
	0.33	0	0.33		
3	0.01	0.98	0.97	1.94	0.97
	0.01	0.01	0		
	0.98	0.01	0.97		
4	0.01	0	0.01	0.04	0.02
	0.01	0	0.01		
	0.98	1	0.02		
5	0.45	1	0.55	1.1	0.55
	0.35	0	0.35		
	0.2	0	0.2		
6	0.45	0	0.45	1.3	0.65
	0.35	1	0.65		
	0.2	0	0.2		
7	0.45	0	0.45	1.6	0.8
	0.35	0	0.35		
	0.2	1	0.8		

FIGURE 2

implement it in R and see how we can compute it relatively easily on a corpus. We need to understand the mathematics of it.

The formula here might look a little bit daunting at first, but I hope to be able to show, and remind you again, that it's actually relatively simple, because remember that it's essentially just computing a number of observed percentages—so how often is the word in question occurring in particular corpus parts—and subtracting from the sizes of the corpus parts also measured in percent. So observed values and expected values, they are both percentages. These are the frequencies of percentage of the word in question in the corpus parts, and these are the sizes of the corpus parts. So then you just compute those pair-wise differences, make them all positive, sum them up, and divide by two.

Again, just to remind you for what follows later, we then look at this in terms of a few examples, which was this spreadsheet. Yesterday I discussed examples one, three and four so let me now discuss number five here, just so that you see one different application. So imagine a corpus that consists of three parts, just like before, and these are the sizes of the parts in the corpus. So one corpus part makes up 45% of the corpus. The second makes up 35% of the corpus,

the third the remaining 20%. So this is really just how big the corpus parts are. In terms of words, if what you're looking at is a word, of course, in terms of constructions, if we had a construction-tagged corpus or something like that. And then the observed frequencies could be where the word shows up how often. And so in this example here, artificial, of course, where we have these three corpus part sizes, here the example stipulates that all of the occurrences, 100%, show up in the largest corpus part and the word that you're interested in does not show up at all, zero, in this corpus part. And it doesn't show up at all, zero, in that corpus part. Then the way you compute this is again, just corpus part size. I mean the difference between this [referring to the second row in Figure 2] and this [referring to the third row in Figure 2] without the minus  $[[0.55]]$ , the difference between this [referring to the second row in Figure 2] and this [referring to the third row in Figure 2] without the minus  $[[0.35]]$ , and the difference between this [referring to the second row in Figure 2] and this [referring to the third row in Figure 2] without the minus  $[[0.2]]$ . So you always just look at how much in percent of the word is in this corpus part minus its size. And then you get these three numbers, you add them up  $[[1.1]]$ , and you divide by two  $[[0.55]]$ .

So what does that mean for our computing dispersion measures for words in a corpus? It means, first of all, that you have to count how frequent is the word in each corpus part. If you have a corpus with ten parts, then for every one of these ten parts, you need to know how frequent the word in question occurs in there. But second, you also need to know how big each of those corpus parts is. So that means if you think about it in terms of planning in advance, it means that as you count the occurrences of the word that you're interested in, at the same time, you should also be separately counting the numbers of words in each corpus part, so the size. So in a sense, you need to do two computations at the same time: how many words are in that corpus part and how many of those words are the words in question?

Yesterday we talked about *enormous* and *staining*, for instance, or *church* and *place*. So if you want to compute the dispersion of *church* in the corpus, you need to know how often does it happen in each corpus part and how large is each of one of those, so that is sort of what we need to be aware of before we start writing any kinds of script, to determine dispersion values for whatever we're interested in. So, for those of you who want to follow along with R, we are now using the script 05 here, `05_dispersion-practice.r`.

For those of you who follow along with the handbook or the program book, of course, you have the html report in there. But so if you want to start this, just double click on it or open it in RStudio or in some text editor or whatever you prefer. It should look a little bit like this, and again with the code up here.

Today, the corpus we'll be using will again be the tagged version of the Brown corpus. First, because it's a very widely used corpus. Second, because it's now available without any copyright issues. And third, because the tagging will make it very easy for us to identify what the words are. We don't need to worry about spaces or any other weird characters—we can just use the tagging scheme as our operationalization of what a word is, and forgetting for the moment that multi-word units like *according to* or *because of*, *in spite of* and stuff like that are not distinguished properly here.

So how do we proceed? Actually, what I've tried to do is I've tried to make this as similar as possible to the things you've seen yesterday. And in a sense, that is really easy because yesterday we were looking at frequencies of words in general, and the word in question are also what we need for dispersion, so much of the stuff that we did yesterday, we will be able to completely recycle here, which of course is nice because it means we don't have to develop the code again from scratch. So some of this will be repetitive or will appear redundant, but hopefully you know it'll just make it clear to you what all these code file commands do and how you can later recycle it for your own work.

So we're going to start actually with the same process as before. So the first thing we need to do is, we need to tell R where the corpus files are. This is using the same code as before so the main function that we used to tell R where a corpus is located is still `dir`, which makes R show you the content of a directory or a folder. And as yesterday, we will call the object that contains all the locations of the corpus files, we will call that `corpus.files`. So what we put into this object is the content of a directory, namely, this one, the one that is called `03_data` and then `Brown_tagged`. That's from the file that you downloaded yesterday. And again, we say `full.names=TRUE`, so we want to make sure that all the path information is actually available. So let me show you that in R. So we run this. So now we just created this object and as yesterday [...].

So we're in a directory that contains all these four script files that we've talked about. That directory also contains this folder: `03_data`. And now we told R, whatever is in `03_data`, and then the directory called `Brown_tagged`, put the names of those files and their locations, put them into this vector `corpus.files`. And so again, you can see we have here the 15 files that all together make up this corpus of one million words of written American English. And so the logic again will be that we want to do kind of the same thing as we did yesterday, just a little bit more than that. So we want to load each file, then we will want to clean it up in some way, namely, we want to get rid of the parts of the annotation that would otherwise would just make things difficult, and then we will want to collect the words. So remember here: This is what that file looks like. So what we will want to get rid of is the line-initial annotation here, and

remember that yesterday we solved that problem by saying ‘delete everything until the first space’, because it’s not a fixed number of characters—because this is one less than here—but it’s always until the first space. And then the way we will want to identify words was we want to say ‘stuff that’s before an underscore, but that stuff must not be a space’. So here, this is before the underscore and is not spaces. This is before the underscore and is not a space, and so on. So those are the two operations we will do here at the same time.

Yesterday, what we did was that we just basically created one super long vector, which was at 1.1 million items that contained all the words from all the corpus files. And the reason why we did that is that, yesterday, we only needed to know how frequent is each word—we didn’t care how often it occurred where. The only thing we wanted to know is, ok, here we have the whole corpus, how often does the word—whatever we looked at—how often does it occur regardless of where. Now, that’s not going to be good enough here, because if we want to compute the dispersion measure, we need to retain the information which corpus part is each word from. And so if you have a vector just of 1.1 million words, all the words in the corpus parts, then it doesn’t say where one file ends and where the next file begins. It’s just all the words in a row cutting across the files and you don’t have any indication of the boundary: This is where Brown\_1A ends. and the next word is from Brown\_1B, and then there’s sixty thousand words or something, and then the next part begins. So somehow we need to retain that information: We need to know for every word: this is from the first corpus part, this is from the second, this is from the third, and so on. So we’ll do everything we did yesterday, but we’ll add one more thing to make sure that we retain that information.

Any idea how one might do that? Well, I guess in terms of implementation, the simplest way to do that would be: If you have an empty vector that will collect all the words, and then what we did yesterday was, we read in the first file, we processed it in some way, and then we put the words into this empty container, so that afterwards it wasn’t empty anymore, it had the words from the first file. And then we added to this, we added the words from the second file, and the third, and so on. So one thing we could do is, we could collect the words from the first file, put them in that container and because we put them there, we know how many it is. We can measure the length of that vector. We can see, the first file contains 50,000 words. And so that means we could create another vector that has the same length and says 50,000, “I’m from the first file. I’m from the first file” like that. So then we have one vector that has all the words from the first file and another vector that says, for every one of these words, which file is it from. And then we load the second file, we process it, clean it up, prepare it, and we add it to the first vector, and then we add for



every second file, we add to the corpus description vector: We add “I’m from the second file, I’m from the second file”, and so on, so that basically it grows at the same time. So instead of collecting one vector with all the words, we are now collecting two vectors: one with all the words, the second one with the information where they’re from, so it seems a little weird, but it means basically we then have two vectors 1.1 million words long. The first one lists all the words: “The Fulton County Grand Jury said ...”, and the other one always says, “I’m from A, I’m from A, I’m from A”, and then 50,000 words later, it’ll say “I’m from B, I’m from B”, and so on. And there’s a reason why this is actually the best way to do it. It’s not the only way, but it’s the best way to do it, especially if you want to compute dispersion. But so this is what we will want to do and the way this could be implemented is, like I said, actually very similar to what we did yesterday.

But the first change is here. So, this time around, we need two collector structures. So this one is what we already did yesterday: We created a vector called `all.corpus.words`. which will collect the whole corpus. So it’ll be 1.1 million words or elements long and so we make that an empty vector. So that’s the proverbial bucket into which we dump words every time when we’re finished with a file. And then we create a second one that is also empty at this point, called `all.corpus.files`. So that’s going to be the vector that says which file each word is from.

So then how do we proceed? Again, we do a loop. And so everything until here is actually pretty much the same as before, but let’s just go over it again in detail. So remember that if you want to do something in R multiple times [...] Like I said, one of the ways to make R do something more than once is a loop, which is this *for*, and then the parenthesized stuff, and then everything that is executed more than once is between the two curly brackets. So the opening curly bracket is here, and the closing one is way down here. So all of this stuff here will be done multiple times. How many times will it be done? Well, 15 times. So again, we say, we create a variable that is called `counter` but of course, you can call it whatever you want. And `counter in 1 to 15` [`counter in 1:15`] says, so, on the first iteration, the first time R goes through all this, `counter` will be 1. Then R arrives down here, goes back up here, and checks “have I been everything”? And it sees I only was 1 right now. So `counter` gets increased to 2, and everything gets run with `counter` being 2. Go up again `counter` is set to 3. Everything is run and so on, until `counter` has been 15, all of this has been run and R checks: have I done this 15 times? The answer is yes and so it concludes the loop [and] is down here and proceeds with whatever is next. And the only reason why we do this is we have 15 corpus files so every one of these corpus files is supposed to be loaded. And remember this is what

happens here: So the square-bracketed notation here says: that of this data structure, the 15 corpus files, always access, first the first, then the second, then the third, and so on. So square bracketing in R is what is called sub-setting, so it selects a subset of values from a data structure.

The simplest way to exemplify this maybe in a different context here would be something like this. So for instance, R has a vector called `letters`, which is just the 26 letters of the alphabet. That's predefined, you wouldn't need to define it. And so if you want to access the fourth letter of the alphabet, you could just say `letters`, and then in square brackets, you say 4 so then you only get that item. If you want the eighth, then obviously, same thing, like this. And so what we're doing in the loop is the same thing: We have a vector of `corpus.files` that looks like this, and on the first iteration, we want to get the first of those. So this is the file that will be loaded and cleaned and processed and that will collect the words. And when it's done, then the counter will be increased to 2. So now we're looking at the second corpus file, and then the third, and then the fourth, and so on, until the last one. So here, every iteration, `counter` gets increased by one, and that means always a new file will be loaded, namely this one, then this one, then this one, until the last one, and we have collected all the words from the corpus. So that's the logic that we're essentially applying here.

Again, we're using the function `scan`. So `current.corpus.file` is the result of loading or scanning of the vector `corpus.files`, the counter-th one, the first one, then the second and the third. What does that contain? It contains character strings, because it's a corpus file—it's not a numeric, a statistical input kind of file. The `corpus.file` is separated by line breaks: We have multiple different lines and then we don't want any output about the sizes of the corpus parts at this point, especially because this would count the number of lines in the file, not what we're interested in, the number of words.

And so after R has arrived here, the first corpus file has been loaded into `current.corpus.file`. So, then as yesterday, we want to clean it up and so what needs to go is this line initial annotation [[for example, SA01:19 the\_AT jury\_NN along\_ RB commented\_VBD ...]]. So again, we'll use this `gsub` function for global substitution: replace: Replace everything from the beginning of a line till you get to see a space. All of that just gets replaced by nothing, which is what does the deletion.

So this is what happens here. So `current.corpus.file` right now is the raw corpus file. And now we're changing it, namely, we're changing it into what you get when you replace everything from the beginning of a line till the first space—I'll explain this in more detail in a moment, but that's what it means: if you replace that by nothing in `current.corpus.file`, `perl=TRUE`, so: use

Perl-compatible regular expressions. This regular expression is actually a relatively simple one. It's a good one to—today we have a little bit more time than yesterday—to explain that in detail. So again, the caret [`^`], that's this roof-like thing at the beginning here, that means the beginning of the line, the beginning of the string. And then, this one here is these three characters belong together. So the period [`.`] means any character: anything, any letter, any number, any punctuation mark, any Chinese letter character, anything but a new line in fact. So this means anything, so that means this period would find a pipe, an “S” and “A”, it can be anything, any one of these characters.

After the period [`.`], you see a quantifier, namely, the asterisk [`*`]. What an asterisk [`*`] means is it says ‘the thing to the left of me, how often that can occur’. And the asterisk [`*`] means ‘zero or more’, which seems totally useless because it would match every time. I mean, if something's not there, it'll match because it's there zero times, if something is there a thousand times, it will also match, because “zero or more”, but there is actually a good reason why one might use something like that. Because, for instance, this is not a good example for this, but what this allows you to do is, for instance, to use one regular expression to find both the British and American spelling of the word *colo*[*u*]*r*. What is the difference between the two? So British is COLOUR and American is COLOR, same with *behavior*[*u*]*r* and a lot of other words like that. And so the way you could find this [...] This is a nice application, so this is a regex editor, essentially. So it's an application that allows you to write regular expression code. [...] So let me show it to you like this. It's short enough to work, so here you see I have the two spellings of the word *color*, the American one COLOR and the British one with the U in here. And so if I now use as a search string, the one without the O, of course, it's only going to find the one without the O, the American one. If I click on “find now” or “evaluate” or something, see then this [*color*] one is found and this one [*colour*] obviously is not. Same if I do it the other way around. If I look for the one with U, it doesn't find this one [*color*], but it does find that one [*colour*].

Now, how do we change this to find both at the same time? This is where the quantifier comes in. Remember, the asterisk [`*`] means ‘zero times or more’, so what we can do is this: We can say the U is optional. So now the asterisk [`*`] means the thing ‘to the left of me, how often’. And so the asterisk [`*`] here says ‘the thing to the left of me is a U that might be there zero times’. And so if it's not there, it finds it. But the asterisk [`*`] also means ‘it could be there one or two or three or four times’ so if it is there, it finds it as well. So that's what the asterisk [`*`] does: ‘the thing to the left of me might be there zero or one or two or a million times’.

So back to this now. So here we're saying from the beginning of the line, there could be anything but it could be there zero or more times. So, in this line, it'll find a pipe [[for example, `SA01`]], which is anything, but it also find the S and the A and a zero and a one, and so on, so all of this stuff will be found. But now the problem is if you only use that, it will actually find the whole line, because the whole line is 'from the beginning and then anything'. So what we need to tell it is, well, 'but stop at the first space' so that it doesn't also stop deleting the words. So that's what the last part does. The question mark `[?]` means 'but only till the first', and then the next character is a space. So, if you follow along, this regex is to be read as 'from the beginning of the line, anything as often as you find it, but only till the first space', that's how that stuff gets taken care of.

Actually, I can show you that here, I mean, put in an example here. So here I now paste a few lines from the Brown corpus into this application as a practice example. And now I wrote the stuff you see in the regex, 'from the beginning of the line stuff till the first space'—you don't see the space here, but it is there. Then I run it and you can see that in every line, it finds the pipe, and the beginning annotation, and then the space, and then it stops. So, it gets rid of what we don't want, because this is what gets replaced by nothing, but it leaves the text with the words alone just like what we wanted it. So that's how this first cleaning-up step works. So the version of `current.corpus.file` that we created in every line will not have this stuff in it.

So then we're here. The next thing we're going to do is we're going to split up to get our words. And just like yesterday, we're going to use the tags for that. And so, again, let me walk you through the expression here but let me also already put this into the other program here. So we begin with `strsplit`—split up the file, which now only has the words and the tags in it—split up that file at this regular expression. And again, I'll tell you what that does in a moment. Using `perl=TRUE`, then we get this data structure that you asked about yesterday. So we get a list, but we want a vector. So when you're done with the splitting, pass this result onto this function [[refers to `unlist`]], which makes it a vector, pass it onto this function [[refers to `tolower`]], which converts it to lowercase, and then put the results, so all the words in lowercase, put them into `current.corpus.words`, so then this thing will contain all the words in the file.

So now how does the regex work? The underscore here is just a tag marker. The underscore would be like here, in the case of *only*, that would be the underscore, and then what does this mean? So the square brackets, what they do is they define what is called a character class. So basically, a list of characters

that you allow to be found or not to be found. The character class that we define here is defined as follows, and this is one annoying thing. Nothing can be done about it, and it's not an R thing. So up here, this caret `[^]` means 'at the beginning of a line'—in square brackets, the caret means something else, unfortunately. You just have to know there's three regular expression characters that have two meanings, depending on where they're used and the caret is one of them. In square brackets, the caret `^` means 'not'. So this character class here, square brackets, and then beginning with the caret `[^ ]` means 'not a space', because the tags are not spaces—they are letters and maybe numbers. So what we're looking for is the underscore, and then this `[^ ]` means something that's not a space, but the square-bracketed stuff fits only a single character, just one. You're saying 'find one character that's not a space', but obviously the tags can be longer than one character. It can be two, can be three—I don't know actually off the top of my head whether they can be four in this corpus—but clearly more than one. So that's what the `+` means, the `+` means 'one or more'. So what this means is: 'find an underscore and one or more things that are not spaces'. So, 'find the underscore and then in this case, two things that are not spaces'. Then it says, 'and after the one or more things that are not spaces, there is a space', and now the question mark means 'zero or one'. So, again, trying to gesture that, which of course is not exactly elegant, but: so the underscore here matches this underscore, then one or more, not a space, matches the R and the B and then the zero or one spaces here matches this space, and then it stops. And that will be used for splitting, which means every tag and the following space will be deleted. And if you look at this here, then that means it leaves us with just the words. This is the word, then the tag and the space that gets deleted, so the word remains. Then there's a word underscore, non-spaces and a space `[[for example, grand_JJ ]]`, this gets deleted. That leaves that word `[[for example, grand]]` so we're stripping out regularly everything that's a tag and any spaces that are still flying around. And then we're done, then we only have the words left.

Let me again show you this here. So this is the regex, underscore, not a space, one or more, till a space. And you can see very nicely here, after every word, it strips out the tag and the space. This is the word—it's not matched—but then the stuff that is matched is the underscore, the tag, and the space afterwards. So all this stuff that's in bold gets thrown out and all the stuff that's in this font here, the light grey smaller font that remains and you can see those other words. This is now what you can properly read as text. "The jury further said in term and presentments ...", so those are the words that are left. So that's how we trim everything down to just words.

Then those get converted to lowercase and stored. And then we do what we did yesterday, namely, we say, the `current.corpus.words` now get put

into this bucket of `all.corpus.words`. `all.corpus.words` now becomes the combination of what it was before, i.e. all previous files if there were any, and now the stuff that you just collected and cleaned. And so yesterday, that's where the loop ended, because we only collected all of the words in one file, in one vector.

But remember, now we said we actually also need to collect where everything is from. So we do something else: We need to, instead of just having a vector of words that is growing with every file, we now also need to say for every word where it's from. And this is what happens here and that's something actually really nice and easy.

So look at what we're doing. We created an empty vector `all.corpus.files`. So at the beginning, it's got nothing in it. And now look what we do: We say now 'make `all.corpus.files` the combination of what it was before and now this'—so what does this do? So `rep` is a function that says 'repeat something'. And I mean, as you might imagine, if you have a function that says 'repeat something', you have to specify minimally two things, namely, what to repeat and how many times. And so what it repeats is the basename of the current corpus file. So right now we loaded `Brown_1_A` and that's what we're repeating and now the magic that happens is that we're repeating it as many times as this file has words. So if the first file turns out to have 50,000 words, then what we're doing now is, we're repeating the name of the first file 50,000 times. If the first file has 100,000 words, we're taking that name of that first file and we repeat it 100,000 times. And so this then is the content of the first file and then the `all.corpus.files` thing will be the same length, but always having the name of that file.

So let's run this and see what happens, not much will happen on the screen. [...] So this is the number of words in the corpus, and this is the number of file names that we have, namely for every word we say which file it's from, so they have to be equally long. Basically, for every word, we created an indicator that says 'this is where I'm from', 'this is where I'm from'. And at some point that will change. So for instance, let's look at the beginning of those two. This is the beginning of the corpus and unsurprisingly, of course, all those are from the first file. This is the end of the corpus and of course, all those are from the last file. Let me check something to show you very quickly: We can now use the function `table`, which tabulates, i.e. it counts, we run the function `table` on this structure that collected for every word where it's from. And so now we can see, the first file has this many words, the second file has that many words, the third file has that many words, and so on.

So because we have this vector `all.corpus.files` that has 1.1 million times all those file names, we just count in this vector, well, how often does each filename occur? The function `table` does that for us. And so, for instance,

we can easily see, this is the smallest corpus part, it seems, only 13,800 words. This one is the largest, nearly 180,000 words, and so on. And so look at this then, so 98,917, so look what we can do with that. 98,917. So let's go with this, 98,920. So this word is from this file, this word is from that file, too, but then the ninth word is suddenly from a different file. So 7, 8, 9, so the first file ends here, and then the next file begins here. That's what we created. Is that clear?

And so the cool thing now is with this, we actually have the expected—we can easily compute the expected percentages. Remember that in that *DP* formula, the expected component was how large are the files, how large are the corpus parts. We have the absolute sizes of the corpus parts here. How do we change that into percentages? No idea? We just divide every file size by the corpus size. We take how many words does the corpus have altogether, and then we divide every corpus part's size by the overall corpus size. So, the first file, that's 8.7% of the corpus, the second file is 5.3% of the corpus, the smallest one is only 1.2% of the corpus, and the biggest one is nearly 16% of the corpus. So this is exp in the *DP* calculation. These are the expected ones. If a word is perfectly evenly distributed in a corpus, then 8.7% of it should be in the corpus file that makes up 8.7% of the corpus. If a word is perfectly evenly distributed, it should be pretty rare in the smallest part. It should be pretty frequent in the largest part. So these are the expected numbers from the corpus part sizes.

So in a moment, we start this again here. So that's how we compute this [[referring to the *DP* formula]], just how big is each corpus part in percent. Now how do we work with that? It's actually relatively simple. I'm gonna skip the graph for a moment. So, one thing you could do now is, you can use this code to very quickly immediately compute for any word that you're interested in the dispersion in the Brown corpus. So how does that work? [...]

So what I'm doing here is I'm saying, now the word of interest is *enormous*, because we looked at it yesterday. It's one of those two words [[*enormous* vs. *staining*]] that occur 37 times in the corpus, but this is the one that was very evenly distributed and *staining*, which we'll look at in a moment, was very unevenly distributed.

So the first thing that I do is I make use of something that in R is very elegant. Let me actually show you that in R, not in the report. So the word in question is defined as *enormous*. And now I'm doing the following. I create a plot. And what am I plotting? I'm plotting this: I'm plotting whether `all.corpus.words` is `word`. What the hell does that mean? Let me show you: Remember this vector? It has all the letters of the alphabet of the Latinate alphabet. So this syntax here, `something == something else`, that does a logical test. Let me show

you what that amounts to. So now I'm saying, here I have a long vector, 26 elements and now I'm saying, show me which one of these things, if any, is C `[[letters=="C"]]`. And the way this works is that R gives you its versions of *yes* and *no*: The A is not a C `[[FALSE]]`, the B is not a C `[[FALSE]]`, the C is a C `[[TRUE]]`, and all the rest of course is not `[[FALSE]]`. And so the cool thing is: this thing can be a million or ten million or a hundred million words long, for every item, it'll tell you *yes* and *no*. This isn't. *No*. This isn't. *Yes*, this is. So what I'm doing here is I'm checking within a split second for 1.1 million words, whether they are the word, which I defined to be *enormous*. And so the response will be 1.1 million TRUEs and FALSEs. Obviously not fun to look at on the screen so we're immediately plotting it into a graph. Now, you might think what do you plot? I mean, TRUE and FALSE—how can TRUEs and FALSEs be plotted? Typically, in Boolean logic of this type FALSE is 0 and TRUE is 1. And so I'm saying `type="h"`, so draw a vertical line that goes from zero to one whenever you find the word. And so this is what that does: It takes a moment because it is 1.1 million comparisons and the plot needs to be drawn.

So here you have that dispersion plot that I showed you yesterday on one of those slides. Believe it or not, but sort of this [the x-axis] is 1.1 million units and whenever the word *enormous* is used, there's a vertical line, and you can see it's distributed relatively unsystematically, relatively evenly in the corpus. It's not everywhere, obviously, because 37 words occurrence is not a lot out of 1.1 million. But it's not like they're all crammed together here, because there's a little bit of it everywhere, so relatively evenly distributed. Let me very quickly, just in comparison, show you what how that plot changes if you look at the other one, namely *staining*.

That's *staining*. All 37 occurrences are in the single location. It's the same number of occurrences as *enormous* in the previous plot, but they're all massed together in a very small space. So that's obviously underdispersed or clumpily or burstily distributed, however you want to call it. So that's how you get these plots basically. Whenever 'is this the word?' is TRUE, then it's not plotted at 0, but at 1 and so those 37 occurrences are all here. But now we also want to compute this dispersion measure so let me go back to *enormous* first.

The old plot here again will show up there in a second. There we go. So now how do we compute the dispersion measure? Remember that we needed to know how frequent is the word in each corpus part, we already know the size of the corpus parts. But now we need to know how often the word occurs in each of them. And again, this is something that in R can be done super elegantly, compared to other programming languages. So what I'm doing here is I am creating an object that is called `wheres.the.word`. That object is a table,



a two dimensional table of frequencies. And so what it cross-tabulates is the names of the corpus files, and whether the words in the corpus are the word in question, *yes* or *no*? TRUE or FALSE?

So now you can see the first corpus part has one occurrence of *enormous* and this many [[98,916]] words that are not *enormous*. The second corpus file has two occurrences of *enormous* and this many [[60,537]] words that are not *enormous*, and so on. And you can see that, in the majority of corpus parts, it's actually included. There's only one, two, and three, it seems, files that do not include it. All the other ones do so that's why it's relatively evenly distributed. So again, this tabulates the names of the files and then whether the word is *enormous*, yes or no? So, but now we need this in percentages, because we need to subtract observed percentages minus expected percentages. So how do we turn this into percentages?

Well, the same thing as before, we would just divide one by the sum of all those, and two by the sum of all those. We take every one of these numbers and divide them by the sum of all these numbers. Then we have a percentage.

So this is what we do here. So the observed percentages are `wheres.the.word`, the column TRUE, this is what this notation means. `wheres.the.word` is this table and this notation means the column called TRUE, divided by the sum of the column TRUE, so 1 is 2.7% of 37 times. 2 is, obviously, then 5.4% of 37 times. So now these are the observed percentages. And the expected percentages actually we can compute from this as well, because where in this table do we see the sizes of the corpus parts? Kind of a trick question, because it's actually not shown. But every corpus part is the sum of this [[FALSE row]] plus that [[TRUE row]]. All the words that are not *enormous* plus here the one that is. All the words that are not *enormous* plus here the two that are. So this table, if you compute the row sums, this [[FALSE row]] plus this [[TRUE row]], this [[FALSE row]] plus this [[TRUE row]], this [[FALSE row]] plus this [[TRUE row]], then you have the corpus part sizes. And then, of course, you can turn them into percentages. And so that's what I'm doing there. Expected percentage is the row sums of that table, divided by the sum of that whole table. So now we have the expected corpus parts like this. And now the only thing that remains is to compute the *DP* value and it is 0.2777, which is definitely closer to 0 than it is to 1. 1 would mean very concentrated distribution, as you will see in a moment—this one is much closer to 0, so it's a relatively even distribution.

So again, the main logic is: collect all the words in the corpus, but in addition to what we did yesterday, also collect for every word where it's from. And the reason why we want to do this is because we want to be able to create this table here, the table called `wheres.the.word`. We need one vector that has all the words in it and one vector that has the corpus files in it, so that we can

generate this, namely for every corpus file is the word in question in there, yes or no? And how many times? And then it's just dividing two sets of percentages and compute that value.

So we saw a *DP* for *enormous* was 0.277, and the plot says that's pretty evenly distributed. Now, let's see what *DP* is, if we apply to the word *staining*. And so the beauty is the whole code now stays the same. All of this is the same as before. The only difference is the first line where we say, now the word we're interested in is *staining*. Everything else remains the same. So if we run this now, it takes a moment, actually, mostly for the plotting, not for the rest, because it is drawing 1.1 million lines ... and done. And so now, obviously, this shows the word is super concentrated in a small part of the corpus. And you can see the *DP* value is now 0.84, so much much higher, much

Yet another example, the two other ones we talked about were *church* and *place*. So again, all this code is the same as before, nothing changes. The only thing that changes is we say now we're interested in *this* word. And from the table here, we can already see this is probably going to be relatively unevenly distributed, because there's a few values that are pretty high. And then a whole bunch of values that are really small. So *DP* for that is 0.5, roughly.

If we do the same thing for *place*, let me scroll down to the value here, then it has a very similar overall frequency. But you can see they are much more evenly distributed. And so this value is only 0.13, showing that is pretty evenly distributed.

And here you can see all the four plots together next to each other. This plot is nice, because it indicates very clearly that these two words [*enormous* vs. *staining*] are equally frequent in the corpus, but they couldn't be more differently distributed. These two words [*church* vs. *place*] are evenly distributed in the corpus. It looks like this [*place*] is ten times more frequent than that [*church*], but it's not. They're actually evenly distributed, equally frequent, but this one [*church*] is just way more concentrated in small sections of the corpus and this one [*place*] is just about everywhere. So that's the reason why frequency may not be as trustworthy as it's supposed to be, because especially here, you know, how much more obvious could the difference be?

And the nice thing now is you can use this code for any other word that you're interested in. Give me any word. Anyone any word that is likely to occur at all in that corpus. *Fault*. So then we just define the word of interest to be *fault* now, and then we run the whole code again. I mean, not the whole code, just this part. It's still plotting. And so there we are: So the *DP* value for *fault* is also about 0.5 and this is how it's distributed. So it's actually relatively specialized: There are huge sections of the corpus where it never shows up, but then here, there's going to be some books or whatever where it's talked about very

much. [...] Take any other word like *atmosphere*. Again, you just run the whole thing, and again, it'll say *church* there, because I didn't change it, but you see the plot will change, 0.29, so actually way more evenly distributed than *fault*, as you can see in quite a lot of places. So however they sampled, they took Library of Congress books and stuff, the word *atmosphere* actually showed up in there a lot of times, and of course this is not sense-dependent. This could be *atmosphere* in the 'meteorological' sense, but also in the 'mood' sense, *There was a bad atmosphere in the room* or something like that. I mean, it doesn't distinguish between them. [...]

So this basically concludes that part. So again, just one more time to remind everyone: So obviously, collect the corpus words, but this time, since we do care where they're from, unlike the frequency list context, also collect where the words show up, because then we want to be able to cross-tabulate the files with whether the word is in there or not. And from that table, you can compute observed, you can compute expected [[percentages]], and therefore you can compute *DP*, and like I said, *DP* actually has been shown in the last two years to perform better on a variety of datasets than the current standard of Juilland's *D*. And so the nice thing is you can run this script on this version of the Brown corpus. It will always work, but the other thing is that if you want to use it on a different corpus, it's very likely, at least, that you will be able to actually recycle most of that script. Because, for instance, if the corpus you're working with doesn't have line numbers, then that actually means you can get rid of this, because you don't need to clean out the line numbers at the beginning. And if the corpus is tagged in a different way, then probably the only thing you need to change in the whole script is this, namely how you use the tags to get the words. So I mean this script with just two changes actually will run on the British National Corpus, will run on a lot of other corpora that have sort of simple annotation of this type. So the idea with a scripting like this, is in fact to be lazy. You want to write the scripts in such a way that, with minimal changes, you can apply it to a completely different dataset. All right, thanks.

## On Association

To recap, basically, I started out the talks here by talking about how frequency on its own might not be the best way to proceed. And in the last theoretical talk, I talked about one additional dimension that I think needs to be considered, which was that of recency, in particular recency, first, in the guise of priming and, second, in the guise of dispersion. So, today I want to talk about the third kind of variable or the third type of dimension that is useful for us when it comes to looking at how things behave in corpora.

### What is the relevance of association/contingency?

- Frequency of form & their dispersion are important, but so is association/contingency (w/ function) – especially for learning, recall Ellis (2006):
- "[l]anguage learning can be viewed as a statistical process requiring the learner to acquire a set of likelihood-weighted associations between constructions & their functional/semantic interpretations"
- association quantifies what-if relations: what [happens] if [the context is like this]?
- "Learning, memory and perception are all affected by frequency, recency, and context of usage: [...] The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization" (Ellis, Römer, & O'Donnell 2016:45f.)
- in other words, association → correlation, → how much does knowing X help you predict Y?
- that's why "human learning is to all intents and purposes perfectly calibrated with normative statistical measures of contingency like  $r$ ,  $\chi^2$  and  $\Delta P$ " (Ellis 2006:7)

FIGURE 1



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611465>

I want to start with sort of bringing back to you one of those several quotes that have been written up by Ellis (2006) and that are very insightful, and how they pinpoint a variety of things that need to come together in a good type of cognitive analysis. For instance, he said, “frequency of form and their dispersion are important, but so is association or contingency with function, and that’s especially true for learning.” Remember this quote, where he said “language learning can be viewed as a statistical process requiring the learner to acquire a set of a likelihood-weighted associations”—the topic of today’s talk—“between constructions and their functional or semantic interpretations”.

The interesting thing or the reason why association is so important is basically that it allows us to quantify *what-if* relations: What happens with some linguistic form, what happens with some linguistic function, if there is a certain context looking like this or like that, or something else? So, pretty much nothing in language happens without any context. We will always be interested in figuring out how the context of something affects either its form (the realization of it in sound or in writing), or its function, its meaning, its pragmatic intention or things like that.

The main other quote then into which we will launch from here again is this one, again you’ve seen it before every time, because it builds up very nicely to what I want to talk about. Again, the quote was that “learning, memory and perception are all affected by frequency” and that was the first theoretical talk; “recency”, that was the second, and now, third “context of usage: [...] The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization.” (Ellis, Römer, & O’Donnell 2016:45f.) So basically, we’re trying to build up, cover, all the aspects that Ellis et al. are discussing in this quote.

So, association basically is concerned with correlation again. Correlation is defined here as “how much does knowing one thing help you predict what something else will be doing? How much does knowing certain linguistic realization of something help you predict its functional impact? How much does the information structure of something help you predict a syntactic realization?” Or something like that. All of these things are what we want to look at.

Again, Ellis (2007:7) already put it very nicely by saying that “human learning is to all intents and purposes perfectly calibrated with these normative statistical measures of contingency [i.e., association like  $r$ ,  $\chi^2$  and  $\Delta P$ ]” and then he actually lists a bunch of correlation coefficients. That essentially is what we want to look at in this talk today: So to what degree does knowing one thing help us predict some other thing where both of these things can be formal or functional realizations of various types of constructions at various levels of granularity or resolution?

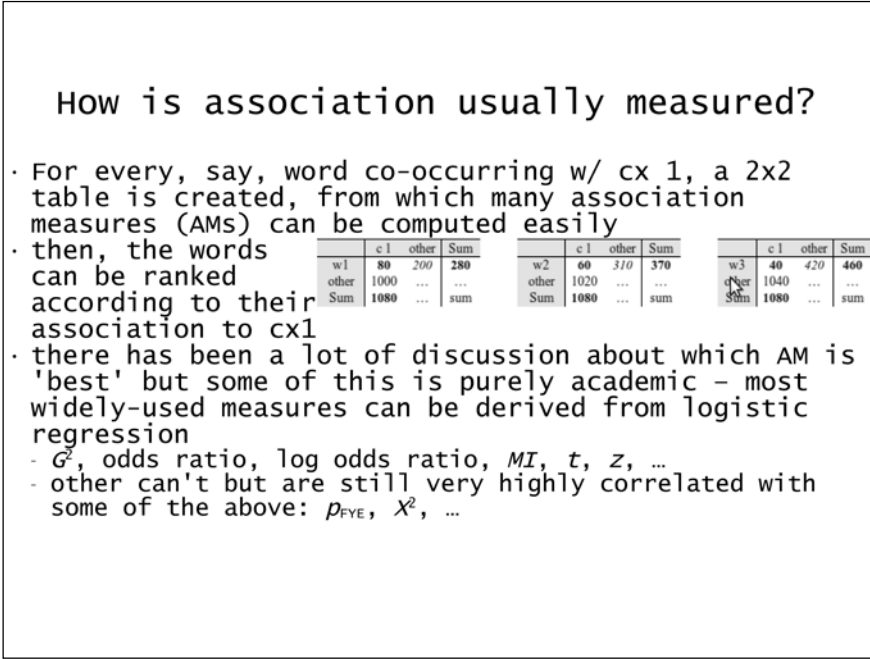


FIGURE 2

Now how is this usually measured? Typically, it proceeds in a way that you've actually seen one time before very briefly in the frequency talk. The example I want to use is that of verb-construction co-occurrence, because so much of my own work has been concerned with things like this. The idea is that for every word, let's say, that occurs with a certain construction, like construction<sub>1</sub> for lack of a better term right now, you draw up a 2x2 table like this [pointing to the first table in Figure 2] from which you can compute association measures. This kind of table here would say basically that word<sub>1</sub> occurs in your corpus 280 times; the construction that you're looking at occurs 1,080 times; and 80 of these two [refers to word<sub>1</sub> and the construction], they co-occur together. The other uses of word<sub>1</sub>: the remaining 200 are with other constructions, not with that one that we're currently looking at. The idea is that you do that for multiple words, ideally for all the words that occur in that construction. So this is the same table, but for word<sub>2</sub> in construction<sub>1</sub>. The construction frequency is the same, and it's still 1,080. But now there's another word that you're interested in, which occurs 370 times in the corpus, but only 60 times with this construction. And here's word<sub>3</sub> occurring 460 times in the corpus, 40 times in the construction, and it's the same construction, so again, 1,080, 1,080, 1,080. Then, when you have an association measure for word<sub>1</sub>, for word<sub>2</sub>, for word<sub>3</sub>,

for each of these  $2 \times 2$  tables, you compute an association measure and then when you have all those measures, you can rank all the words by their association to that construction. As we've seen in a ton of literature that has been using this collostructional approach, a lot of times you'll find that the words that like to occur in certain constructions, they share certain semantic characteristics, sometimes they share certain information-structural characteristics, and other kinds of things and so they allow us to interpret semantics of constructions for instance.

One big issue of discussion for many years now has been, in corpus linguistics actually for decades, is what association measure to use. There's a lot of different measures that can be applied to a deceptively simple  $2 \times 2$  table like this. Many of you may have heard of a chi-squared test, for instance, as a test that is routinely used for tables like this. But the chi-squared test, for instance, makes some assumptions that a lot of times are violated with these kinds of data. So people have come up with literally many dozens of association measures that can be used to quantify the strength of association in these tables.

As I say here, much of the discussion, however, is actually purely academic because many of the measures that are being used most of the time in corpus linguistic approaches are all just different ways of interpreting logistic regression results. So a logistic regression is a regression that tries to predict [or] model something binary, namely this construction in question (construction<sub>1</sub>) or another one (other constructions) on the basis of a binary predictor, namely this word (word<sub>1</sub>) or another one (other words). So actually, it would look like overkill, but you can, instead of doing a chi-squared test on a table like this, you can do a logistic regression on a table like this and the results will be typically at least very, very similar.

So once you adopt this regression-modeling perspective, then actually many of the different measures that people have been hitting each other over their head with are all very, very comparable. For instance, the most frequently-used measure probably is this one,  $G^2$  (which some of you may know as the log-likelihood ratio), odds ratio, and log odds ratio: All of these values actually come from a logistic regression. If people sort of debate whether using this or this is better, they're actually not debating very much, in the sense that these will all be extremely highly correlated.

Some other measures cannot be derived from logistic regression, like the ones that I mentioned here, Fisher-Yates exact test ( $p_{\text{FYE}}$ ), the test that a lot of people have been using in collostructional kind of analyses, or chi-squared. But they're still extremely, highly correlated with anything that comes out of a logistic regression. If you, in general, are interested in looking into this kind of issue more—because association is an important concept in usage-based

## How should association be measured?

- The following considerations are relevant to choosing an AM
  - **symmetry**: is the AM supposed to be symmetric or not?
    - nearly all AMs are:  $p_{FYE}$ ,  $LLR$ ,  $\chi^2$ ,  $MI$ ,  $t$ ,  $z$ , log odds ratio ...
    - some are not:  $p(y|x)$ ,  $\Delta P$ , ...
  - **metric type**: +effect -freq. vs +effect +freq
    - the former: log odds ratio, the asymmetric ones above, ...
    - the latter:  $p_{FYE}$ ,  $LLR$ ,  $\chi^2$ , ...
  - **frequency information**: token vs token+type frequency
    - the former: all but one
    - the latter: lexical gravity  $G$
- probably best settings in an ideal world:
  - symmetry: no
  - metric type: +effect
  - (frequency: token+type)
- ideally **dispersion** would be included in some way
- let me suggest two measures for your consideration
  - log odds ratio
  - $\Delta P$

FIGURE 3

linguistics—then learning something about binary logistic regression is probably time well spent, because it will help you understand all sorts of debates and all sorts of results that have been published on these kinds of questions.

Now, if this is how association usually measured, then how *should* it be measured? There's a bunch of characteristics is that you should pretty much always consider when you talk about, or when you consider, which association measure you think is best for your particular case study. The first one is this, namely the question of symmetry. Nearly all association measures are symmetric. All the ones that are listed here, Fisher-Yates exact test, log-likelihood ratio, chi-squared, Mutual Information, all these statistics basically are symmetric and by *symmetric*, I mean they quantify how much a verb and construction are attracted *to each other*. Another way of using symmetry or describing symmetry here would be that the association is bidirectional: The word likes the construction, and the construction likes the word to the same degree. That's what is meant by *symmetry* here.

But there are some measures ( $p(y|x)$ ,  $\Delta P$ ,...) that are not symmetric, so that means these measures here. This would be a conditional probability: What is the probability of this construction given this verb? That would be different from what is the probability of this verb given this construction. So with these



measures, you *can* distinguish between cases where a verb likes a construction a lot, but the construction doesn't like the verb a lot. You can keep those things apart. That's probably a useful thing, because it's not really obvious at all that associations that we form in our minds as part of a learning process would be symmetric. Usually, if only temporarily, we see something first, and then we see something else so chances are that that has at least some kind of impact on the degree [and direction] of association we form between these things.

The second important characteristic, and that one has been debated particularly hotly in collostructional analysis literature, namely, is the type of metric (+effect -freq. vs. +effect +freq.) that your association measure is. There's essentially two options, to simplify a little bit here. One of the two metrics is this, namely, the association measure reflects association (+effect), but it does not reflect frequency (+effect -freq.). Whereas the association measures that are mostly used are of the latter type, so they reflect association strength, but also frequency (+effect +freq.). So in a way, or one other way to look at it would be that, some metrics measure only one dimension, namely, how strong is the association and I don't care in how many data points I observe this. Other metrics measure an association, but also take into consideration the sample size, the number of items you have, and give you that back in one number. So measures that do not include frequency would be something like the log odds ratio or the asymmetric measures that I've mentioned here, conditional probabilities and  $\Delta P$ . The most widely used ones, actually, like log-likelihood ratio, Fisher-Yates exact test, they react to both the association strength and the frequency with which something has been observed.

It's still an ongoing debate which of these two scenarios is better. I'll talk a little bit about what I'm thinking, but just to give you a heads-up already, this one [pointing to the latter type] is simpler to use, because for every word and construction pairing, [[for example,]] for *give* in ditransitive, for *tell* in ditransitive or something, it gives you one value and that value reflects both effect and frequency: It's easy if you're statistics-averse and you want just one little value to sort by.

However, like I've already indicated the other day when we talked about dispersion, there I said, conflating of frequency and dispersion into one adjusted-frequency value loses a lot of information. That, of course, happens here as well. The reason why one might consider something like this [pointing to the former type], a measure that only measures association strength but not frequency, is to keep your data clean. The statistic that you're reporting only looks at effect size, because frequency, you have that anyway, so that would be the second axis in a plot. That's something we'll look into later.

Then, third: frequency information. Pretty much all of the association measures that are widely used only use token frequency. That means, these  $2 \times 2$  tables that you've seen before, you don't know how many *different* other constructions a word shows up in.

Let me go back a real quick to show you what I mean here. In this case [[Figure 2]], we know word<sub>1</sub> shows up 80 times with this construction, and we know it shows up 200 times with other constructions—but you don't know how many different constructions these 200 other ones are—could be 1 or could be 200, but we don't know. Pretty much all association measures but one that I know at least work like that: they just take the 200 and they do not consider how many other competing constructions are there. The only measure that does use type frequency as well is a measure that is hardly ever used, namely, lexical gravity *G*. Computing it is a little bit more involved, but theoretically, of course, it seems like a very useful idea in fact.

Now the best settings ideally would be, probably, to use a measure that is not symmetric so that you can distinguish cases from where the verb likes the construction, but the construction does not like the verb, or the other way around.

Second, probably, at least for cognitively, supposedly, realistic analysis, you probably want to use only an effect size [[+effect]] here, so that your results for association are not tainted also by frequency, but you keep those two things separate. Frequency, ideally, one would be able to use both token and type frequency, although no one has done that yet, for reasons that may become apparent later.

Then, ideally, you would also include dispersion because we've already seen that a co-occurrence frequencies like these can be very misleading depending on how high the dispersion is. If this corpus has 20 parts and all the 80 co-occurrences are in one of those 20 parts, whatever you're doing with that, you should know that, as opposed to these 80 being distributed all over the corpus being very representative. So, ideally, we would include dispersion here as well and I'll show you a little bit about how to do this at a later point. Now, obviously, if there are many dozens of association measures—the last overview paper that I've seen discussed 80—and since then at least one or two others have been developed as well, so there's more than 80 of those, what should you be using? Let me suggest two here for your first consideration at least. The first one is the log odds ratio, and the second one is  $\Delta P$ . Let me now tell you why I think these two are useful and should be considered.

The log odds ratio is a symmetric measure. So, that's already kind of a downside. But it has another good thing, and that is that it only is a measure of association and it does not reflect frequency so it's cleaner than something like

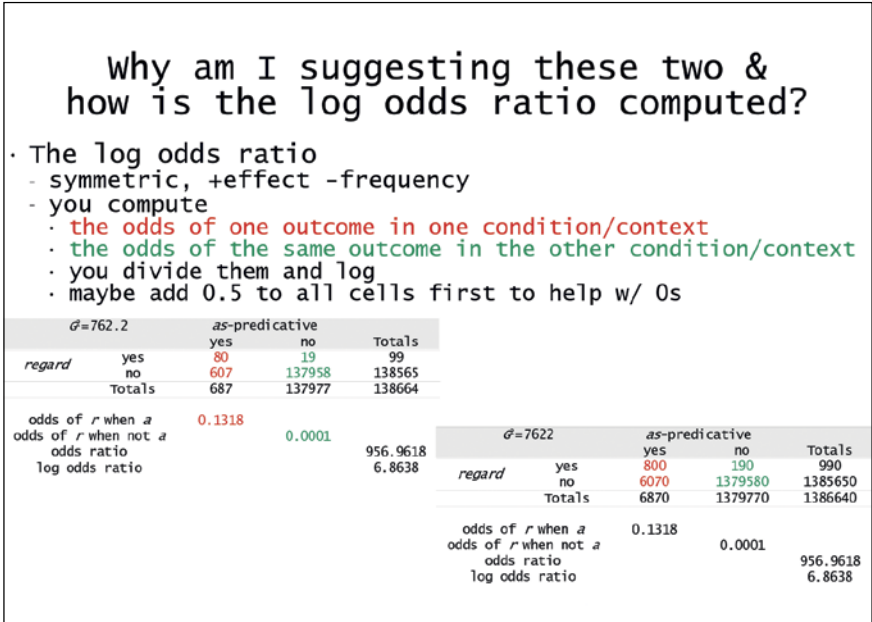


FIGURE 4

log-likelihood ratio or something like that. So how's it computed? Let's use this construction and this verb as an example. We're again looking at the as-predicative construction. So that is this construction, what did I say, whatever, *He was regarded as a very famous linguist*, that would be an example. *He saw himself as a very important linguist*, *He described himself as a very important linguist*, *He considered himself as a very important linguist*" this kind of construction. So, verb, a direct object, *as* and then something. *This attack was widely regarded as being out of the blue*, that would be another example.

So we're looking at this construction, either that (*regard* in as-predicative) it's there or it's not, yes or no, and we're looking at this verb *regard*, which is either there or not. In the corpus that we're looking at, the corpus contains this many verbs. That's what we're using as a unit of sampling here. Of those verbs, 99 cases are *regard*. And of those 99 cases of *regard*, 80 are in this construction. So most, the vast majority, I mean 80 percent, pretty much, and 19 uses of *regard* are not in that construction. The construction has a frequency of 687, and 80 of those, so one eighth essentially kind of, are with *regard*, and then there are 607 others.

So how do you compute the log odds from this? I don't know how many of you bet on horses or something, but it's basically that type of odds. So you divide the number of times that something happens of interest, this number

here [referring to 80], by the other option that [referring to 607]. So 80 divided by 607 is 0.1318. These are the odds of *regard* being used when the construction is an *as*-predicative. There are 687 *as*-predicatives, and the odds of *regard* are this (*regard* in *as*-predicative) versus this (*regard* not in *as*-predicative). This many cases of *regard* “yes”, compared to this many cases of *regard* “no”, when the construction is in fact an *as*-predicative. So those (0.1318) are the odds of *regard* when the construction is an *as*-predicative.

Then you compute the same thing here: what are the odds of *regard* when the construction is not *as*-predicative? And as you can see they’re tiny: 19 divided by 137,958. I stopped here, I rounded it off at four decimals, but obviously it’s very small. So then the odds ratio is this (0.1318) divided by that (19/137,958), which gives you this number (956.961) and then you log it. That’s the log odds ratio (6.8638), and this is a pretty damn high value on that scale. Again: what are the odds of *regard* versus not when it’s the construction of interest divided by what are the odds of *regard* compared to it’s not when it’s all other constructions, and then this divided by this, log—so relatively straightforward. You can do this with any spreadsheet, even if you wanted to do it with a pocket calculator, not particularly tricky. Sometimes, what you need to do is, if one of these numbers is zero, or in cases, some of these numbers are zero that you add 0.5 to every number first, and then you do the computation that I showed here.

The important thing here to realize: so, first, it’s symmetric. This is the number that says how much the verb *regard* and the construction *as*-predicative like each other. Second, like I said, this does not include frequency information. This is counter to other measures. Here, actually, I’m not showing you how this is calculated, but this ( $G^2 = 762.2$ ) is the log likelihood value for this table, 762.2 is also super high. Now, what happens if we pretend we had a corpus ten times as big as the one that we’re using here? We are multiplying every one of these numbers by ten. What happens then is this: The odds ratio, the log odds, all stay the same. If this is ten times larger, and you divide it by this, which is now ten times larger, of course, you get the same number. But look at this ( $G^2 = 762.2$ ). That value ( $G^2 = 762.2$ ) went up by a factor of ten. So, this value doesn’t separate corpus size and association—it conflates them into one value, whereas this one (log odds ratio = 6.8638) nicely only includes the association strength. So that’s one of the potential selling points for this measure.

Then, what about the other one,  $\Delta P$ ? The thing about  $\Delta P$  is that it also does not grow if the corpus becomes bigger. Just like the log odds ratio, that’s what they have in common but  $\Delta P$  is asymmetric: So  $\Delta P$  can distinguish how much the verb likes the construction from how much the construction likes the verb. That’s why here, I’m writing how is  $\Delta P$  and then the  $c \rightarrow r$  means column/row.

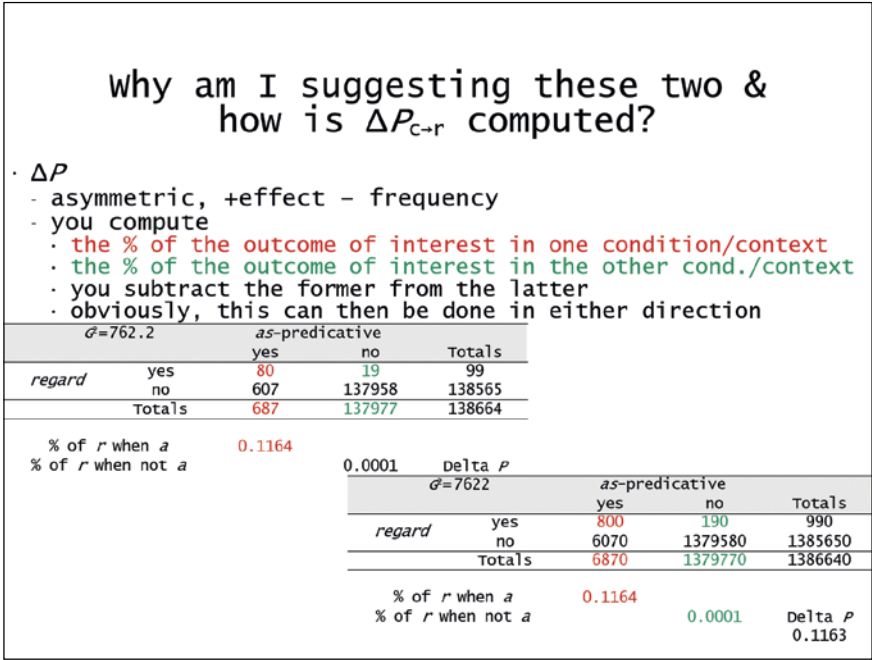


FIGURE 5

So how much does  $\Delta P$  from whatever is in the columns, the construction, like whatever is in the rows, the verb, how is that ( $\Delta P_{c \rightarrow r}$ ) computed? We're looking at the same example. And it's actually kind of similar. So this is the same table as before.

What you compute is you compute the percentage of the outcome of interest in one condition. If there's an *as*-predicative, 687 times, how often in percent is that *regard*? We're computing how much is 80 out of 687? So it's 0.1164. This is really just saying 11.6% of the *as*-predicatives are with *regard*. Then we do the same: how often is a verb *regard* when it's *not* in the *as*-predicative? That's, of course, super rare. There's a buttload of verbs in general, but only 19 of those are *regard* because in general it's not a frequent word. Then you just subtract this (0.1164) minus this (0.0001), and that (0.1163) is the value. What this tells you is, how much does knowing that the construction is the *as*-predicative help you expect *regard*? When the construction is not the *as*-predicative, *regard* is super rare. But when the construction becomes the *as*-predicative, it's quite common, and  $\Delta P$  is the difference between those two. So it ranges from minus one to plus one and the higher, the stronger the attraction.

Here we're going from columns to rows. We're predicting from the absence or the presence of a construction which verb is going to happen. Here too, this

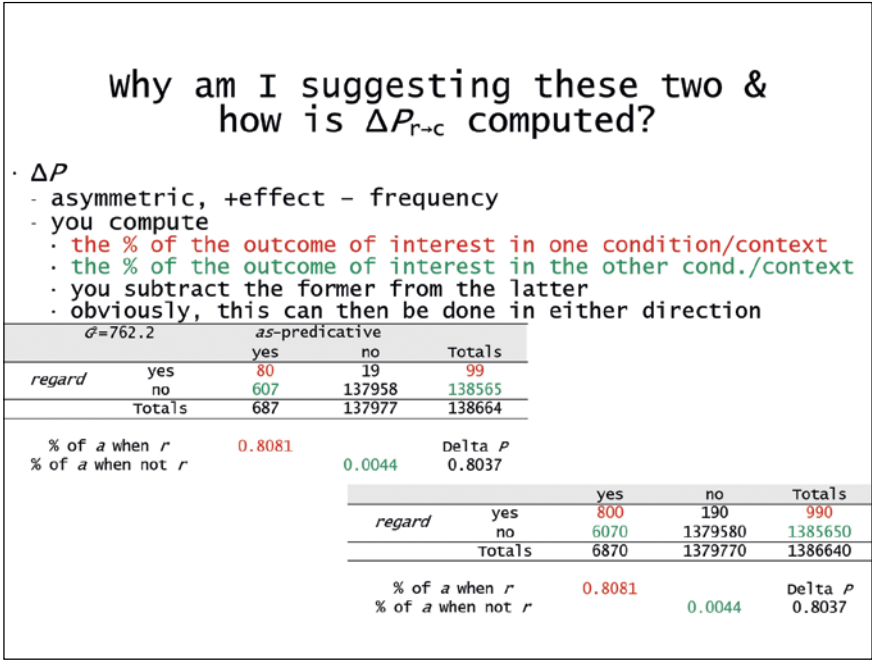


FIGURE 6

thing does not reflect frequency: If we multiply that whole table by ten again,  $\Delta P$  is the same. So, [[it is]] very nice, keeping things separate. This is the example for from the construction to the verb.

Obviously, since this is asymmetric, we now also need to look at from the verb to the construction. Somewhat confusingly, but there was no other way to show it in a parallel way in a spreadsheet. So now we're saying if the verb is *regard*, how often do we see *as*-predicative? This (80) out of this (99) is a staggering 80%. If you see *regard*, you can be pretty certain, it's in an *as*-predicative.

But then the other one is how often do you see an *as*-predicative if the verb is not *regard*? Not very much. Again, the  $\Delta P$  value is the difference between the two. You can see this one (0.8037) is super high. That's what  $\Delta P$  or any asymmetric measure buys you. This measure here can 'see' that, well, if the construction doesn't actually attract a verb that much, but the verb attracts the construction super strongly.

I'll come back to this a little bit later in a moment. But, since it fits here right now, let me mention it already. There's a lot of cases where corpus linguists and psycholinguists alike, they're very annoyed at the fact that sometimes corpus data don't match up nicely with experimental psycholinguistic data. But, for instance, I've seen one example, someone used corpus data and tried to

correlate association measures from corpus data with the result from an association experiment, sort of, given one word, and then someone was supposed to give words that they associate with that word. Then the corpus-linguistic author basically correlated the psycholinguistic results with corpus associations of this type. But what she did is, she used a bi-directional association measure when of course the experimental task was totally directional. Namely, you get one word and you're supposed to go from that verb to somewhere else. So part of the mismatch of course could very well be that she used an association measure that actually is not compatible with the experimental task to which she is trying to compare the corpus data.

Same thing with the sentence completion task: If you do a sentence completion experiment, and the sentence fragment ends in a verb and then you look at how do people complete that sentence. That's a directional question. You give a verb and you expect to see a construction after that. So you need this: how much does a verb boost the appearance of a construction? Not the other way around, and not something symmetric.

That's one of these cases where psycholinguists and cognitive linguists always like, "well, corpus data, they are not really that great" and then they use their wrong measures and you're like, "yeah, of course they're not that great, if you don't do the math right". Here, too, again, this does not change, if the corpus size gets much, much bigger. So, quite an attractive measure.

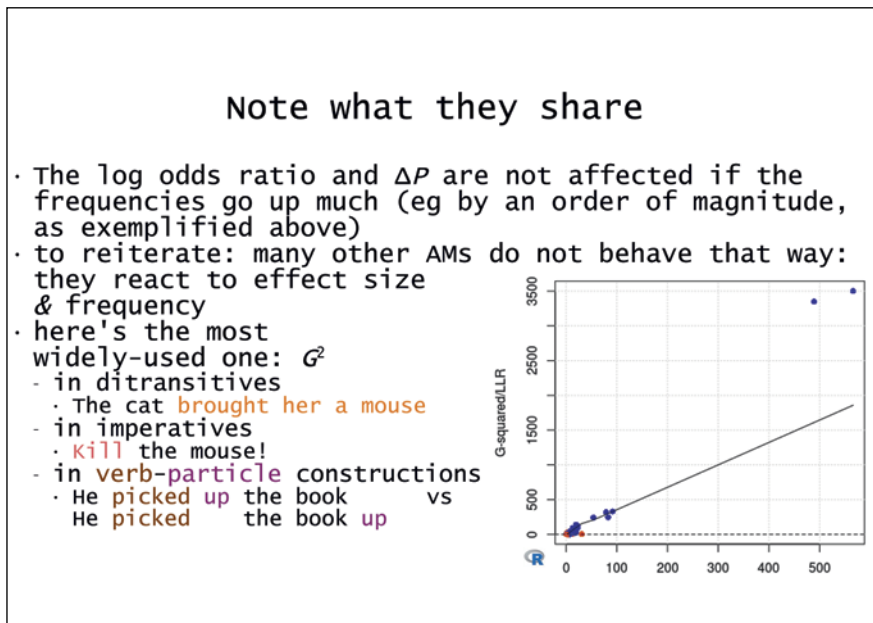


FIGURE 7

So what do they share? Again to recap, both of them are not affected by the corpus size increasing a lot, but actually increasing at all. If everything is as before, then the measures will stay the same. And that is not how most measures react: Many standardly-used measures, they do incorporate both. Again, I think there is an area of application for that: If you're interested in a simple-sorting kind of result, but for anything that aims to be cognitively realistic, I think this is not the way to go.

So let me show you three examples using the most widely used association measure, the log likelihood ratio. Here's an example: We're looking at the ditransitive constructions, like *The cat brought her a mouse*. This plot here, what it shows is the frequency with which a verb occurs in that slot and the association measure, so every blue or red point is one verb in that construction and you can see there's a relatively strong correlation. The regression line doesn't capture quite those here but you can see that on the whole, there's an upward trend: As the frequency goes up on the whole, the points go higher up, even if most of the values are clustered down here, because of the Zipfian distribution.

Same thing with imperatives, so which verbs go into this slot (*Kill the mouse!*)? Again, there are some outliers, but on the whole, there's a positive relation as frequency increases, so does this association measure.

Finally, for this one, that's another example that we'll discuss later, verb-particle constructions. So the two constructions in question are *picked up the*

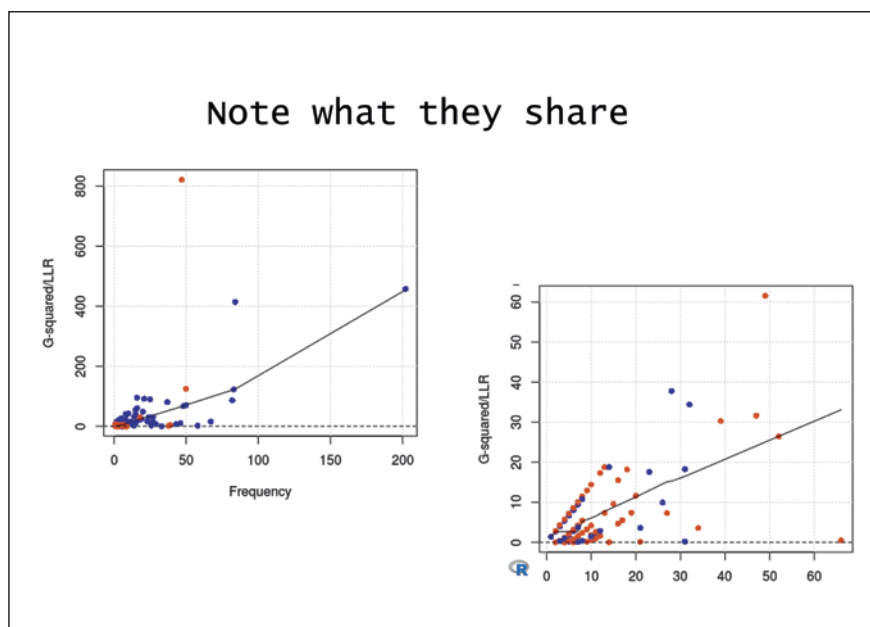


FIGURE 8



### Note what they don't share

- The log odds ratio is symmetric,  $\Delta P$  is not, ie
  - the former cannot distinguish these collocations,
  - the latter can
    - *of-course, at-least, for-instance, in-vitro, de-facto, ...*
    - *according-to, upside-down, instead-of, ipso-facto, ...*
    - *Sinn↔Fein, bona↔fide, ...*
- in the spoken part of the BNC, all of these have
  - $G^2 > 178$
  - log odds ratio  $> 5$
- but why would such learned connections would be (as) symmetric? (Trautschold 1883, Cattell 1887)
- in fact, mismatches between corpus and psycho-linguistic data might be in part due to overlooking the directionality of collocations

FIGURE 9

*mouse* and *pick the mouse up*. Again, there's a very clear correlation between the association measure and frequency here. So this is potentially a problem: if you're interested in keeping those dimensions separate, then this measure does not do that. Whatever you measure here, to a large extent, it actually also reflects this, you're not keeping them separate.

Now, what do they *not* share? Like I said, the log odds ratio is symmetric so it is somewhat less informative if you want to put it that way because the log odds ratio will not be able to distinguish certain kinds of collocations from each other. These are all collocations where the second word is highly predictive of the first. If I ask you what word might you expect in front of the word *instance*, of course you're going to say *for*. If I ask you what word you might expect in front of *least*, it's very likely that you would say *at*, maybe you would say *the*, *the least I can do* or something. Here this one is particularly nice, in front of *facto*, what's there going to be other than *de*?

But there are collocations where it's the other way around. I mentioned this example earlier, I think. If I ask you what comes after *according*, of course, you're going to say *to*; if I ask you what's after *instead*, of course, you're going to say *of*. But if I ask you what's in front of *of*, chances are you give me a whole bunch of different things as well so that correlation is not that strong. And so here we

## But is $\Delta P$ really worth it?

- Given how  $\Delta P$  is computed, it is
  - correlated much w/ transitional probability  $p(x|y)$
  - only natural to ask whether it's different enough from  $p(x|y)$  to even make a difference
- Schneider (to appear): yes
  - data: Switchboard NXT 2008 (642 phone conversations)
  - dependent variable: hesitation placement in PPs
  - predictors:  $a$ ,  $\Delta P_{\rightarrow}$ ,  $TP_{\rightarrow}$ ,  $\Delta P_{\leftarrow}$ ,  $TP_{\leftarrow}$ ,  $MI$ , lex. grav.  $G$
  - statistical analysis: party::cforest
  - results: many different results for the three kinds of PPs, but
    - "it is mostly  $\Delta P$  which outperforms transitional probability"
    - this is true for both forward-directed measures and backwards at phrase boundaries
    - $\leftarrow$  measures are good predictors of collocation status when  $w1$  = function word &  $w2$  = content word
    - other major finding: lexical gravity  $G$  does very well!
- Dunn (2018): tuples of different  $\Delta P$ s are useful

FIGURE 10

have *facto* again, actually" if you start from *facto*, then you end up at *de*. But *facto* is also the thing that people would say occurs after *ipso*. And then there are cases which are completely predictable in both directions, at least in some corpora, *bona*↔*fide* or *Sinn*↔*Fein*, perfectly predict each other. And so the log odds ratio will treat them all the same. It would not distinguish between them whereas  $\Delta P$  would establish these three groups.

Like I said, log odds ratio for all of those is greater than 5, pretty high. But there's no reason to assume that these kinds of associations are, in fact, symmetric because, like I said, if only because of time, there's always going to be one thing first, and then the other thing second. Then we might interpret correlation or co-occurrence but it's not obvious at all that it would be symmetric.

Now one question you might have though is whether the effect of  $\Delta P$  is actually worth it. Do we really need to compute this? Because given how  $\Delta P$  is computed, it's extremely highly correlated with transitional probability.

Let me actually show you that in the table again.  $\Delta P$  is this (800) divided by that (990), that's one probability, minus this (6,070) divided by that (1,385,650), that's another. So obviously, these two (0.8081 and 0.0044) are transitional probabilities. And  $\Delta P$  (0.8037) will be always be very highly correlated with this one (0.8081). Why? Because usually, the  $d$  cell here, which is not this construction

and not this verb, is usually very high, just like here (i.e. 1,379,580). If you look at one word and one construction, then of course most of the corpus is going to be other things. So that number (i.e. 1,379,580) is always going to be very high, which makes this (i.e. 6070) divided by that (i.e. 1,385,650) very small. So  $\Delta P$  will be very highly related to this (i.e. 0.8081). And so there have been people actually who suggested that just take this number (i.e. 0.8081), forget about the normalization, you know, minus this (i.e. 0.0044) to that (i.e. 0.8037).

So the question might be, is it even worth it to compute it like that? Maybe we can do it without it. But there have been some first studies now that show that it's worth it. So Schneider's (2018) book, as a part of her dissertation, did a study where she looked at data from the Switchboard corpus so phone conversations between strangers put together on a switchboard. She looked at the hesitation placement in prepositional phrases. Where do people become slower because they're entering into an area where there's production planning difficulties? And she compared a whole bunch of different predictors, co-occurrence frequency,  $\Delta P$  in one direction, the transitional probability in that same direction,  $\Delta P$  in the other direction, the transition probability in the other direction, other collocation measures, and so on, and then she did a random forest analysis on this dataset, really nice.

She did find a lot of different kinds of results for three different kinds of differently complex prepositional phrases but one of her main conclusions is this, namely, "it is mostly  $\Delta P$  which outperforms transitional probability"; "both for forward-directed measures and backward-directed ones". So yes, it does come with some work, but there will be enough cases to make it worth your while, which is essentially what she's finding.

Then in another paper, I'm not sure it's still to appear. Just this morning I saw a reference to it on ResearchGate, and it's said at 2018, so a study in the *International Journal of Corpus Linguistics* where James Dunn look at a whole bunch of different  $\Delta P$  values and he also found that it is an extremely useful concept. So yes, mathematically, there will be a high correlation, but that should not detract from the fact that, on the whole, the higher degree of precision of  $\Delta P$  is better. If you're in the market for an association measure, so to speak, then  $\Delta P$  is probably a pretty good one.

Now how this might be applied? This is where I do want to talk a little bit about collostructional analysis, even though it's kind of dated by now, but still a lot of people are using it. It's a method that, like everything else in corpus linguistics, is based on the distributional hypothesis, which I'm giving here again, "If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions from A

## Collostructional analysis

- Collostructional analysis (CA) is an method based on the maybe most fundamental corpus linguistic assumptions: the distributional hypothesis
  - "[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution" (Harris 1970:785f.)
- CA is a straightforward extension of ...
  - of collocations: co-occurrence of words/lexical units
  - to (one sense of) colligation: co-occurrence
    - of words
    - and patterns/constructions

FIGURE 11

and B are more different than the distributions of A and C'. So, in a sense, it's a straightforward extension of collocation work in corpus linguistics as it has been happening for decades. The only difference or the main difference being that instead of looking at words co-occurring together, we're looking at the co-occurrence of words in, or with, patterns or constructions.

Three different methods have been distinguished. The first one would be essentially the type that you've seen before. The first method is collexeme analysis. You're looking at one construction, which is in the columns here, construction<sub>1</sub> (yes versus no); construction<sub>1</sub> (yes versus no) and you're looking at a bunch of words, each of which occurs at least once in the construction. The construction<sub>1</sub> occurs 1,080 times, word<sub>1</sub> occurs in it 80 times, word<sub>2</sub> occurs in it 60 times, and so on. The idea is, for every word, you compute a measure of association: for this one, for that one, and for all other ones, kind of like what we discussed before.

The second possibility, maybe actually even more widely used, because it's simpler, is distinctive collexeme analysis. You have two, or theoretically more competing constructions—competing in the sense of they are functionally similar. For instance, they might constitute one of those famous argument

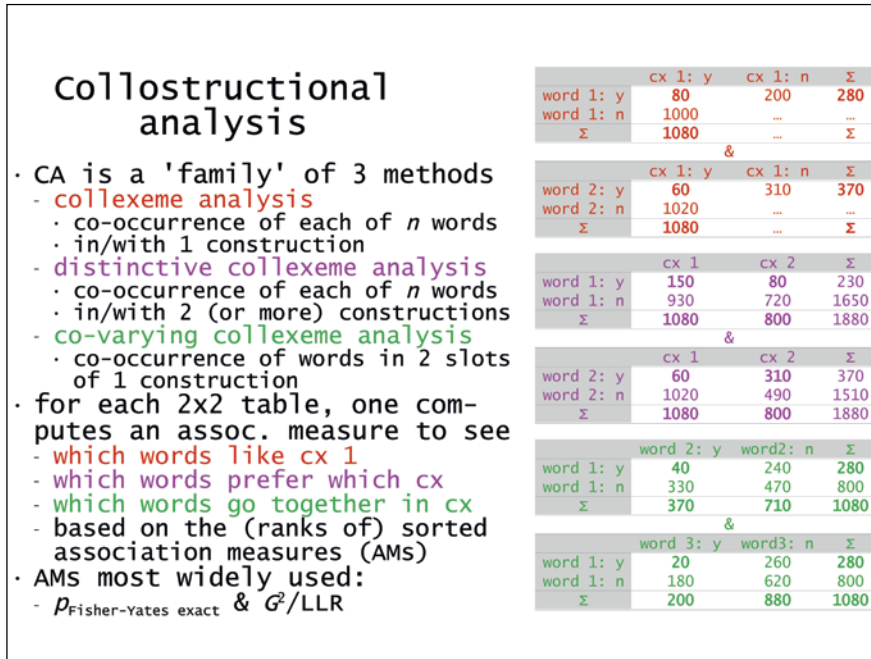


FIGURE 12

structure alternations, or something like, like ditransitive vs. prepositional dative or something like that. So you find every instance of this construction, every instance of that construction, and then every word that shows up at least once in one of the two. So this would be a case where word<sub>1</sub> strongly prefers to occur in this construction (150) as opposed to that one (80). Whereas here we have a word<sub>2</sub> that very, very strongly prefers to occur in this one (310) as opposed to the other one (60). And so distinctive collexeme analysis would quantify that and would compare the two with each other.

Finally [co-varying collexeme analysis], not used that much, although it's also interesting sometimes. You have one construction with two slots in it, and you're looking at co-occurrences sort of depending on what happens in the first slot, what's going to happen in the other one. So it's word<sub>1</sub>, yes or no, and word<sub>2</sub>, yes or no, in the same construction and then you can quantify the preferences there.

If we want to do this, but also address some of the problems that I've mentioned before—so the fact that something like log-likelihood conflates frequency and effect size, the fact that all association measures do not take dispersion into consideration—then how can we do that? We're going to look at a few examples where we try to address at least some of these things. So we

## Addressing at least some of the above-mentioned problems

- Let's look at a few examples of CA, where we
  - keep frequency and contingency separate
    - using (log2) of the observed co-occurrence frequency of verbs & a construction
    - using an association measure that doesn't include the observed co-occurrence frequency
  - add dispersion to the mix
    - computing the dispersion of, say, verbs in the construction against the distribution of verbs in general
- examples
  - collexeme analysis: ditransitive
  - collexeme analysis: imperative
  - distinctive collexeme anal.: verb-particle constructions
- things not to be discussed (much) here:
  - keeping directions of association separate
    - we could use  $\Delta Ps$  as AMS ( $\Delta Pv_{fromc}$  &  $\Delta Pc_{fromv}$ )
  - no entropy
  - no polysemy

FIGURE 13

will keep frequency and contingency or association separate. We're not going to use a measure that grows even if just a corpus size grows—we will use a measure that only grows if the effect becomes, in fact, stronger.

So the example that I'm going to use here is that for frequency, we're going to look at the log frequency, because in psycholinguistics research, most of the time, we find frequency effects on a log scale. Then, as an association measure, we're going to use the log odds ratio for now simply because we already have a variety of dimensions to juggle and I don't want to add two association measures: verb to construction, construction to verb, to the mix at the same time.

Then we'll add dispersion to the mix by looking at how evenly are the instances of the verb in the construction attested throughout the corpus. We want to avoid this example that I talked about yesterday: We want to avoid cases where verbs like *fold* or *process* score high on association strength, although they only show up in a construction in a single file—that's what we want to protect ourselves against.

We're going to look at three examples, namely the three constructions I showed you before on the slide (Figure 12). For log-likelihood ratio, [[first,]] we're going to do a quick look at a collexeme analysis of the ditransitive,

obviously one of the most widely studied constructions out there. Second, we're going to look at the collexeme analysis of the imperative because that's a construction where yesterday we saw dispersion causes problems so we'll now going to check, can we handle that? Third, we're going to look at a distinctive collexeme analysis case, so do we find verbs that prefer the order where the particle comes before the verb and are there verbs that prefer the other order where the particle comes later? So are there verbs that prefer the construction, *He brought back the mouse*, and are there other verbs that prefer the construction, *He brought the mouse back*? In fact, you will see that there are quite strong tendencies.

Now we're not going to look at here at this point, simply because we don't have the time and the complexity quickly becomes quite daunting, we're not going to look at different directions of associations, and we're not going to look at entropy or polysemy at this point in time. I have data for at least this part and that part—polysemy, I haven't done yet myself. That awaits future research. Let's build this up step-wise.

We're going to first look at the ditransitive construction. I wrote a small script that basically gives us the number of ditransitives in the British component of the International Corpus of English. We find a pretty Zipfian distribution here as always: The 1,820 ditransitives, that's 88 different verbs showing up

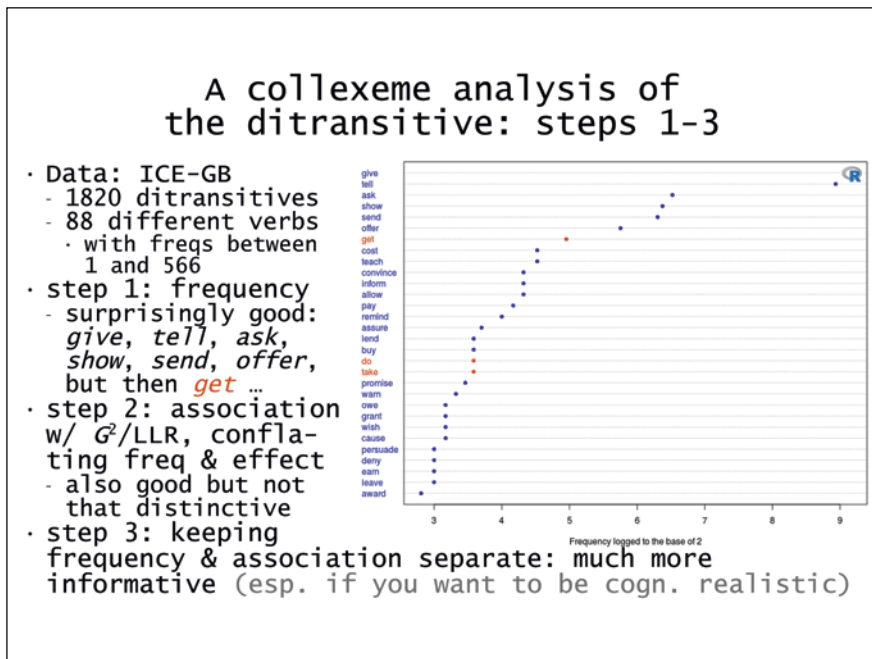


FIGURE 14

in that construction and they have frequencies between 1 and 566. One verb actually is like nearly like 30% or something of all the instances but then there are also quite a few cases that show up only a single time.

Now, if you look at the frequencies with which verbs show up in that construction, then you can plot it like this [referring to the plot in Figure 14]. Here on the  $x$ -axis, we have the frequency, but again, it is logged, but you can see, *give* is the most frequent word and here is a dot that's behind this [pointing to the dot that represents the frequency of *give*]. That's the 566. Then *tell* is a little bit less frequent. Then there's a huge gap already, everything else is way below that. As before, actually, the first results are quite good, given the semantics that one, after decades of researching this construction to death, after decades of this, if you look at the verbs, *give*, *tell*, *ask*, *show*, *send*, *offer*, that's exactly what everyone has always been talking about in that construction. No big surprises there.

Again, I do want to point out though the verbs you see here in red, *get*, *do* and *take*, especially *get* shows up in the construction quite frequently, but actually less often than expected: *Get* is an extremely frequent word in general so while this is a high number, it's actually too low a number because given how frequent *get* is in general, you know, this number should be way higher than it is. Again, the ranking by frequency alone doesn't even tell you.

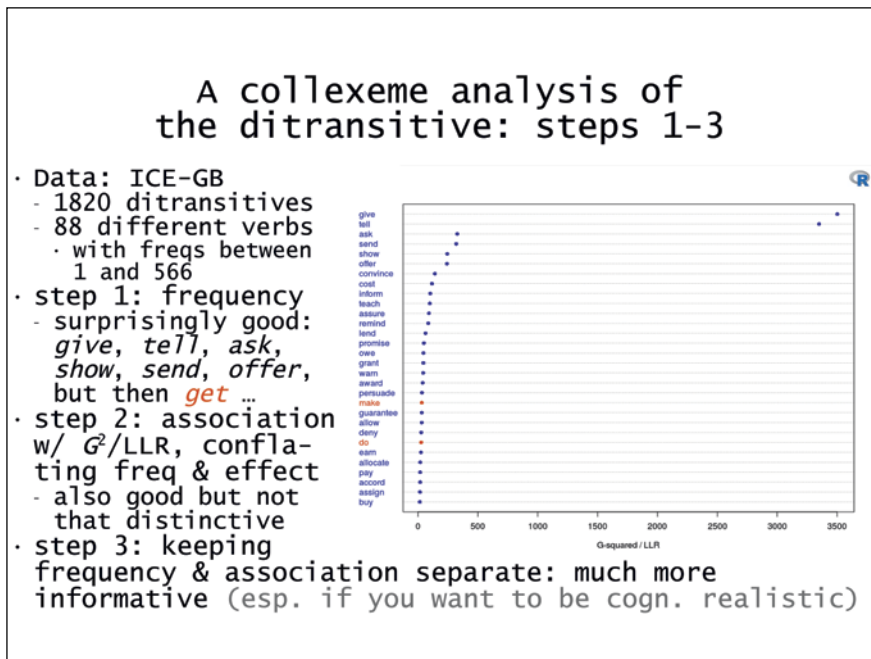


FIGURE 15



Now let's compute an association measure. This time around, actually first, the one that is not that great, namely, the likelihood-ratio, because it conflates frequency and effect. You can see, again, *give* and *tell* win out, but this time by a huge margin. Because now, on top of the fact that they already lead in terms of frequency [referring to the graph in Figure 14], this distance is even made bigger because of association coming to the mix. So now the verbs that are repelled by the construction actually have very, very low associations. It's better in that sense—we see which words are repelled by the construction—but it's worse in the sense that there's a clear conflation of frequency and effect size.

Now what about keeping frequency and association separate? This would be one way to do this and it's actually kind of interesting for a reason, I'll discuss in a moment. So we have a frequency on the *x*-axis, again logged. This is 4, 16, 64 and so on [referring to the frequency represented on the *x*-axis]. You can see *give* is more frequent in the ditransitive than *tell*, [because] it's to the right of *tell* but *tell* is higher up. So the association of *tell* to the ditransitive is higher than that for *give*.

That's something that the normal kind of measure doesn't tell you. It doesn't tell you where a certain value comes from, whether it's the more frequent component or whether it's the more attraction component. Here you can see that

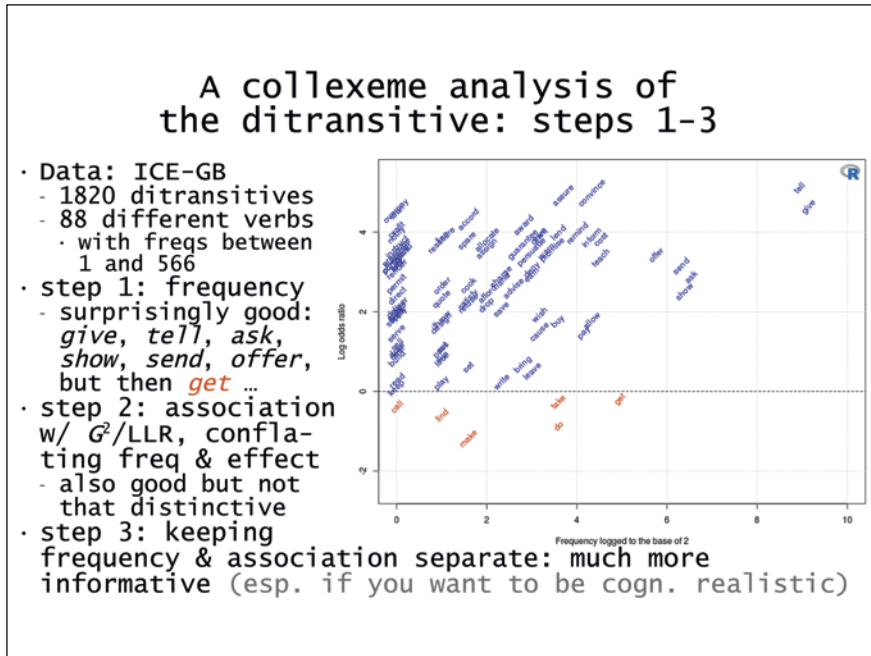


FIGURE 16

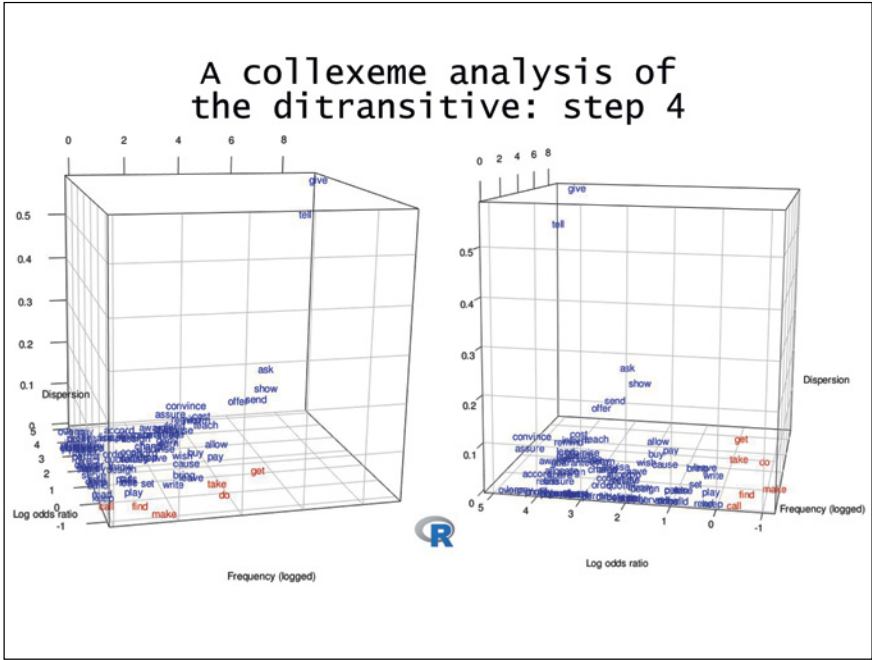


FIGURE 17

very well. You can see for all the words that are repelled, their log odds ratio is negative. And you can see up here that the correlation between frequency and associations are actually not that strong. It's not like we have a nice point cloud that goes up like this [referring to a linear relation] but we kind of have a pretty big mess here with maybe something going up slightly. What this clearly shows is that our keeping frequency and association separate works. You can't look at this axis [x-axis] here and completely clearly predict what's going to happen on this axis [y-axis]. Frequency and association are not the same. For cognitively *realistic* analyses, there's no good motivation to conflate them into one number, pretending they are the same and hoping for the best.

Now this [Figure 17] is what happens when we add dispersion. This one is a little bit hard to interpret. I am going to show you the interactive version of this plot. [[The explanation of the 3-D plot]] This axis here is frequency on a log scale. You see the frequency values here. It's always opposite of the legend: This is the label for this axis, and the axis is shown on the opposite side of the cube up and top. Then this is association, the log odds ratio. The red verbs are all the ones that are repelled, the blue words are all the ones that are attracted. The further to the right, the stronger the attraction. Then this axis is dispersion.



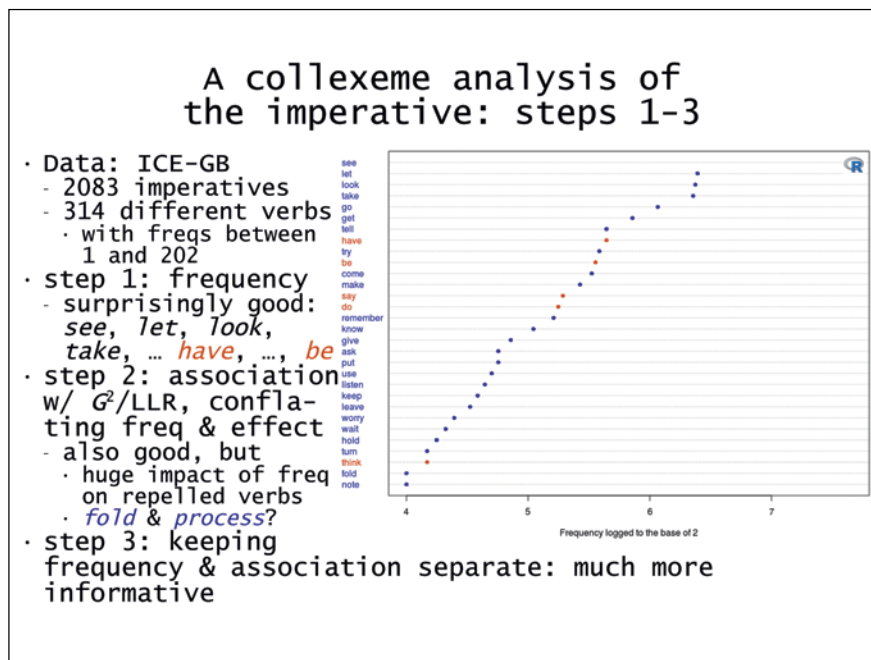


FIGURE 19

Look at this point here where the mouse is right now, that point is the origin of that cube, that's when every dimension is zero. It's when this dimension is zero, it's when this one is zero (you are at the bottom of the cube), and it's when this edge is zero. So if you want one number, IF, what you could do is, you can take that cube, so keep your eye on that point here, and so you measure the distance from the origin to where the verb is. You have a cube like that with the origin here, if a verb is located down here, up here, and then in the back, then that is the Euclidean distance from the origin to that point. And if you do it like this, that number will capture frequency, recency, and association. Again, there's a problem, which I'm not going to bother you with now, but theoretically, that's one way in which this could be done.

I think we at least agree that this is way more informative than putting it all in one number or just reporting a frequency.

What about the same for the imperative? This is the example where we hope that our approach can help deal with *fold* and *process*. We have about 2,083 imperatives, about 314 different verbs in them, with again, Zipfian-distributed frequencies between 1—hapaxes—and 202. If we sort it by frequency, again, it looks like this [referring to Figure 19]. Examples like, *see* and *let* and *look*, again it's up here, pretty good. But again, also, *have*, *be*, *say* and *do* are quite frequent

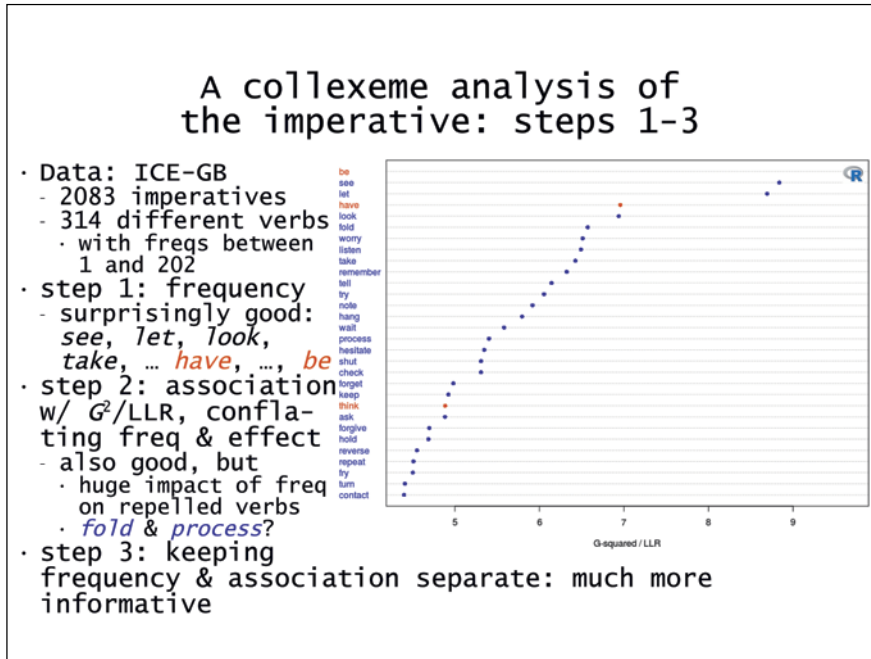


FIGURE 20

in the imperative, but less often than you would think given how frequent they are in general. But the frequency doesn't show you that. You only see that because I marked it with the color. If I didn't put the color in there, you wouldn't know that this is actually repelled by the imperative. And it's obvious, you don't often see *have*, *have* what, it's not a verb that lends itself for the imperative.

Here's the log-likelihood value. This one actually is really funny, because the highest log-likelihood value is scored for a verb that is repelled by the construction. You would have to make that negative actually so that it gets sorted at the bottom. But then we get *see*, *let*, *look*, and *fold*, there the problem verb is and here is *process*, the other problem verb that we hope a bigger approach, the better one, that will now not rank that highly. Again, *have* is very high up but actually repelled, the same with *think*, you don't use *think* in an imperative very much.

Again, what we want to do is first to keep frequency and associations separate. We have frequency here now. And so *see* wins very strongly, quite a bit of a gap till the next one. But actually, as you can see, the association is actually not that high—there's a ton of verbs that are more strongly attracted to the imperative. That means the fact that *see* here wins—it's the highest blue

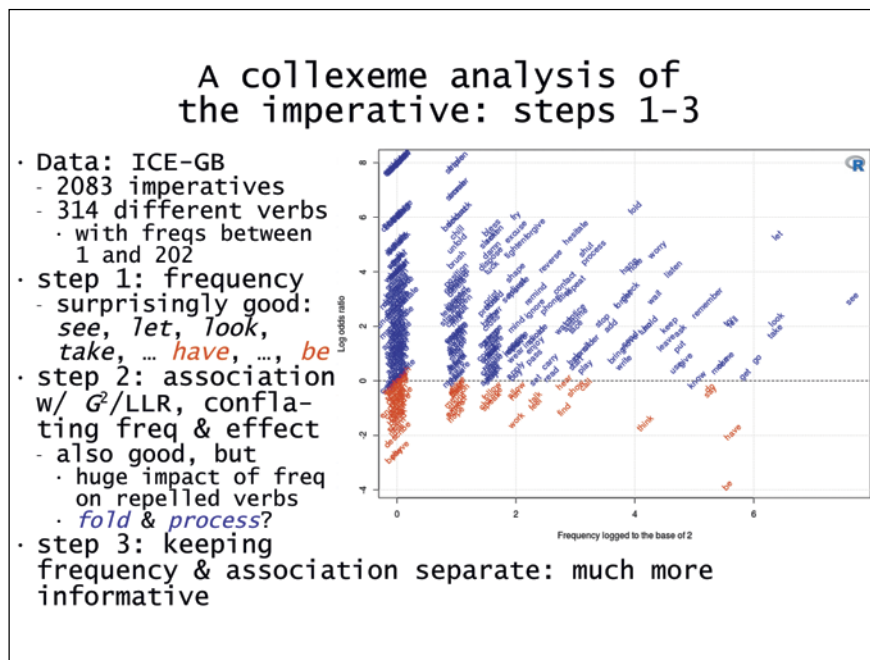


FIGURE 21

verb—the fact that *see* wins here, that's not due to a super high association strength, it's due to a higher frequency, which again, from looking at this value, you don't know that. You only know that if you keep them separate and see, what does it do here, what does it do here? Things like *let*, *worry*, *listen*, and *hesitate*, *don't hesitate* probably, *forgive* as in *forgive me* probably, all those have a higher degree of attraction to the imperative than this one. But actually, so does *fold*. Not that frequent, but very strongly attracted to the imperative. Even more than *hesitate* and *shut* (which is probably *shut up*).

Again, we want to keep it separate. We're getting this [referring to a 3-D plot similar to the one in Figure 22]. So here we have the same thing. Frequency is on this axis, log odds ratio on this axis, and dispersion on this one. The verbs that are [[blue are attracted]], the words that are shown in [[red are repelled]], if you apply a *post hoc* correction. And we can see, *see* seems to be the overall winner. How does that happen? First, now this is frequency, so it's more frequent than everything else. Secondly, it's not more attracted to it than everything else, but it's way more evenly dispersed in the imperative than everything else, that's why it has this marked position here up at the top.





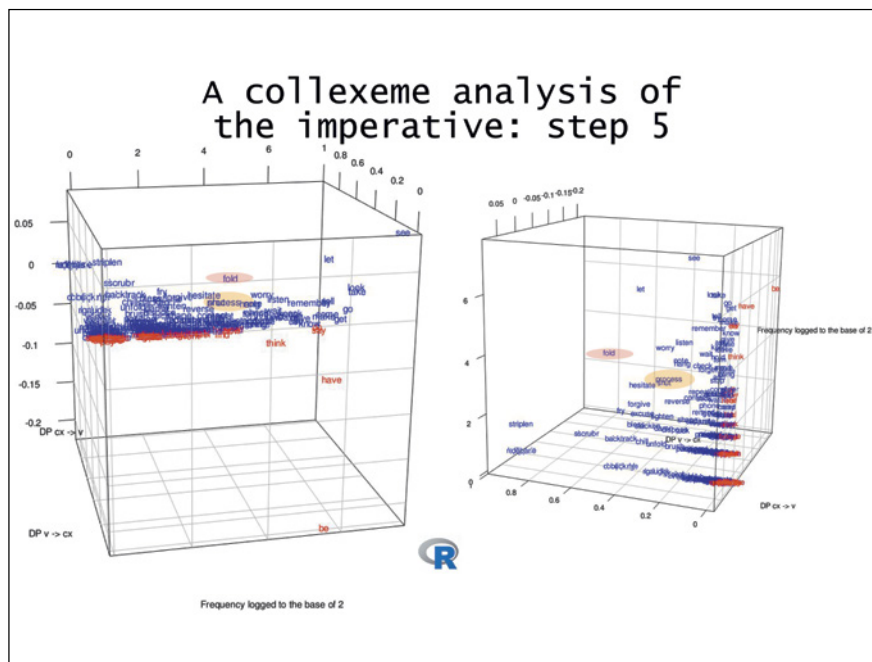


FIGURE 23

the verb to the construction in this axis, the one going into the back of the wall; and then construction to verb, that's the one going up. You can see, for instance, here's *see* and here's *let*. *See* is still scoring high on frequency, but here with this attraction, it's close to zero. So what *see* does with that construction is that—the construction likes *see*, but not the other way around. You know where there's a ton of verbs where it is the other way around. So here then we have *fold* and *process*, the ones whose dispersion puts down and downgraded, so that we didn't make a wrong conclusion there.

So, final example, maybe speeding this up a little bit in terms of time: We're looking at the verb-particle constructions so this is a distinctive collexeme analysis, the results look a little bit different. This dot chart now is organized from left to right. The verbs that score highly are the ones that like verb-direct object-particle, verbs like *put out*, *get up*, *put up*, *bring up*, they like the construction where the particle is at the end. And verbs here at the bottom with the negative values, they like the other order, namely, verb-particle-direct object. So, *carry out*, *set up*. These data say that you're more likely to say *pick up a book* than *pick a book up*. So, here in the middle, I omitted a whole bunch of verbs that are not distinctive, just to save space on the plot.



## A distinctive collexeme analysis of verb-particle constructions: steps 1-3

- Data: ICE-GB
  - 1164 VPCs
  - 835 different verbs
    - with freqs between 1 and 31
- step 1: frequency
  - hard to evaluate, seems reasonable(?)
- step 2: association w/  $G^2$ /LLR, conflating freq & effect
  - seems very similar w/ ranking changes
- step 3: keeping frequency & association separate
  - much more informative

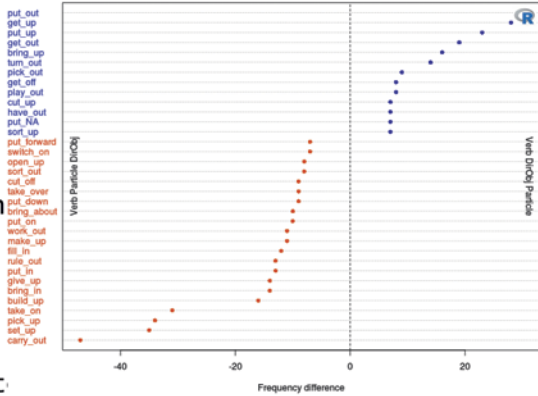


FIGURE 24

## A distinctive collexeme analysis of verb-particle constructions: steps 1-3

- Data: ICE-GB
  - 1164 VPCs
  - 835 different verbs
    - with freqs between 1 and 31
- step 1: frequency
  - hard to evaluate, seems reasonable(?)
- step 2: association w/  $G^2$ /LLR, conflating freq & effect
  - seems very similar w/ ranking changes
- step 3: keeping frequency & association separate
  - much more informative

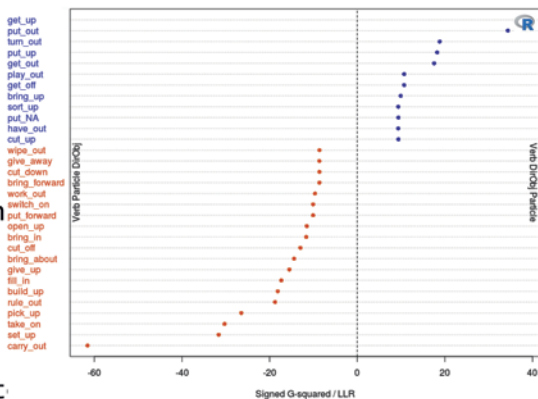


FIGURE 25

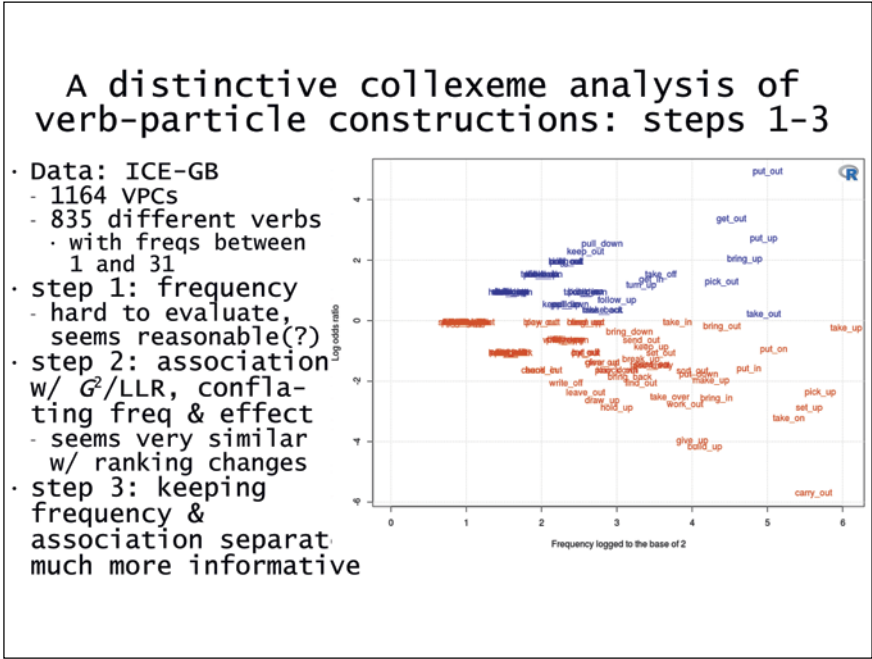


FIGURE 26

It doesn't change much actually, if you apply the likelihood-ratio test here, as the results are really very similar. We still have *get up*, *put out*, *put up*, *get out* here at the top, and we still have *carry out* here at the bottom, so not much of a change. That's in part of course because this is so much affected by frequency as well.

But if we look at frequency and associations separately, we get a very different picture: For instance, some of the most frequent verbs, the verbs on the right, actually have no strong attraction at all. They go equally well with both. *Take up* has no preference for either construction, but something like *put out* or *carry out*, they have very strong attractions, that's why they're very high up in the plot and very much further down in the plot. And again, not really a correlation here between frequency and association, but it's not like you would draw a line through this, and it would be clearly going down or something like that.

Then this would be the three dimensional representation again. Frequency is the back axis, this is association, either for this construction or for that construction, and this is dispersion. Here, those are the verbs that are most strongly attracted to this construction (verbs attracted to VPO are colored in red) and this is the verb that is most strongly attracted to that construction (verbs attracted to VOP are colored in blue).

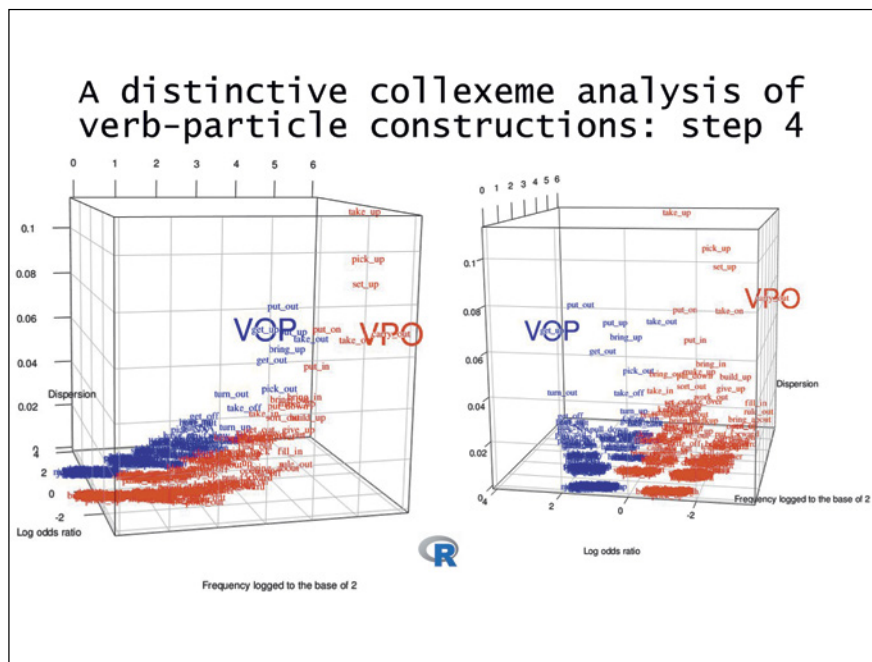


FIGURE 27

I'm not going to push it quite that far. But it is really, really tempting to use the word *prototype* somehow in that connection. Because in terms of frequency and association and dispersion, this is the verb that most goes with that. Again, I'm not going to push it quite that far, but it's tempting to at least think about that.

To wrap up, collostructional analysis as a method in general has been very widely used, no doubt about that. There have been diachronic studies, there have been synchronic studies, this method has been used in first and second, or foreign language acquisition, it has been used successfully in studies having to do with priming effects for native and non-native speakers, actually.

The exact implementation varies between applications. Not all applications use the same association measure and it's not obvious always which association measure to use. I give you two recommendations but depending on what exactly you have in mind for your study you might have different ideas about that. But the more important point is this, namely, that the logic of including association per se is sound. We can debate how we measure it but given the psycholinguistic, the psychological, and all sorts of linguistic literature itself, we do want to make sure that this is a dimension of information we do include. That means, you shouldn't believe all sorts of nonsense that you can

Concluding remarks re collostructions  
(from Gries 2012, 2015)

- Collostructional analysis has been widely applied
  - diachronic & synchronic construction studies
  - first & second/foreign language acquisition
  - psycholinguistic studies of priming, ...
- while its implementation may need to vary between applications, the association logic per se is sound
- so don't believe all sorts of nonsense about it
  - no, the use of AMS – *p*-based or otherwise – is not a big significance testing problem but maybe a conflation one
    - conflation of effect & frequency
    - conflation of direction of association
  - no, the other-other cell (d) is not a huge problem – you estimate it reasonably
  - no, semantics doesn't go *into* it, but it might *emerge from* it
  - so, if you criticize it for something
    - you better understand it first
    - provide alternative measures that are as good or better
    - *then* we can talk ...

		as-predicative	
		yes	no
regard	yes	80 (a)	19 (b)
	no	607 (c)	137958 (d)

FIGURE 28

read about collostructional analysis in some not-to-be-named-here publications. For example, some people have harped on the fact that values like log likelihood or *p*Fisher-Yates exact test, that's like a huge problem because of all the null hypothesis significance testing issues that you run into—that's not really necessarily the case. If anything, the problem is a conflation one, namely, that you conflate frequency and effect size but the fact that the measure is based on the *p*-value per se can, in fact, be even corrected for.

Some other people have talked about how difficult it is to compute this number (cell *d*), all the instances that are not the verb in question and that are not the construction question. And again, at this time, I'm not even saying who said this, but this is just nonsense. You estimate that number on the basis of everything else that you have here. If you're looking at a construction and a *verb* slot, then obviously this number will not be determined by the number of *nouns* in the corpus, but by the number of *verbs*. Plus, if you do simulations, it doesn't even matter that much how high that number is so don't believe that part, like computing this cell is so difficult that you can run this.

Someone has criticized the analysis for that it disregards semantics. That is true—but only in the most trivial sense. The point is, semantics, that doesn't *go into* the analysis—it *comes out of* it. So once you've done the rankings of all

## Concluding remarks re association

- In terms of learning, acquisition, & processing, there's little that's more important than association
- association measures quantify
  - what-if?
  - if ..., then ...
- different measures are available,
  - all based on frequency of occurrence and co-occurrence
  - but differing in terms of implementation & implications
  - which shows that 'frequency' per se is versatile, if used properly and non-anxiously
- thus and not forgetting all previous 'lessons'
  - include frequencies of occurrence & co-occurrence
  - be aware of direction of association
  - be aware of dispersion
  - be aware of whether you can or cannot tolerate the information loss resulting from conflation - if yes, conflate properly

FIGURE 29

the verbs in that construction, then typically at least you can interpret that in a semantic way. The idea of this is to *prepare* you for a semantic analysis and not to *presuppose* one. In other words, if you criticize the method, then you'd better understand it first, and provide measures that are as good as the ones that are being used to better. Then obviously, we can talk. I mean, at least some of these claims are just demonstrably false actually, even in the papers that criticize the method for it.

Then last slide. I would go so far as to kind of support Nick Ellis and colleagues here very much: In terms of learning, acquisition and processing, there is really little that's more important than association, because nothing happens without a context. Everything will be tied to some condition in some way. What association measures do is they quantify basically *what-if* or *if ..., then ...* scenarios: 'what is going to happen if this is the case?', or 'if this happens, then what will be the next corollary of that?'.

There's a whole bunch of different measures that are available, minimally 80. They all are based on frequency of occurrence and co-occurrence but they're different in terms of how exactly these frequencies are used and that also means they're different in terms of what you can take away from them. Again that means that frequency is a relatively versatile notion that goes way

beyond just how often does something happened, *if* you use it properly and not in fear of any imperialists. So to not forget all the previous lessons, do look at frequencies of occurrence, do look at frequencies of co-occurrence, but then also be aware of direction of association: Don't always assume by default that something is bidirectional. Be aware of dispersion, you've seen in a 3-D plots how much of a difference it can make. Be very careful with conflating things into numbers, because a lot of times that information loss might kill exactly what you're interested in. And now this thing crashed. Thanks.

## Association: Practice with R

To get started for today, you will need the next one of these script files, `07_association-practice`. My recommendation would be that you just go to this folder, which contains all the stuff that you downloaded from the website. Double click on this file so that RStudio will open. What you need to see is the script file on the association measures example that we're going to deal with today.

If you're following along with the handbook, obviously you will have this file. Like I said yesterday, what I want to talk about most here is the importance of association, so how do we measure to what degree two things are associated with each other? Either in the sense that something attracts something else or in the sense that something repels something else—where a certain construction is, a certain verb may be not seen that very much.

In order to use an example here that uses a verb- or a word-construction association, we're going to switch corpora. So we're going to look at a different type of corpus here. And the corpus we're looking at is the one that in your folder is in the folder called `o3_data` and then find `ICEGB_sampled`. This folder contains data from the British Component of the International Corpus of English. This is a corpus that is tagged and parsed and manually corrected, so it's a very, very reliable source of syntactic information. However, that corpus, as far as I know, is not freely available so what I did here is I basically trimmed it down a little bit so that you only have the words, but you don't have much of the syntactic annotation.

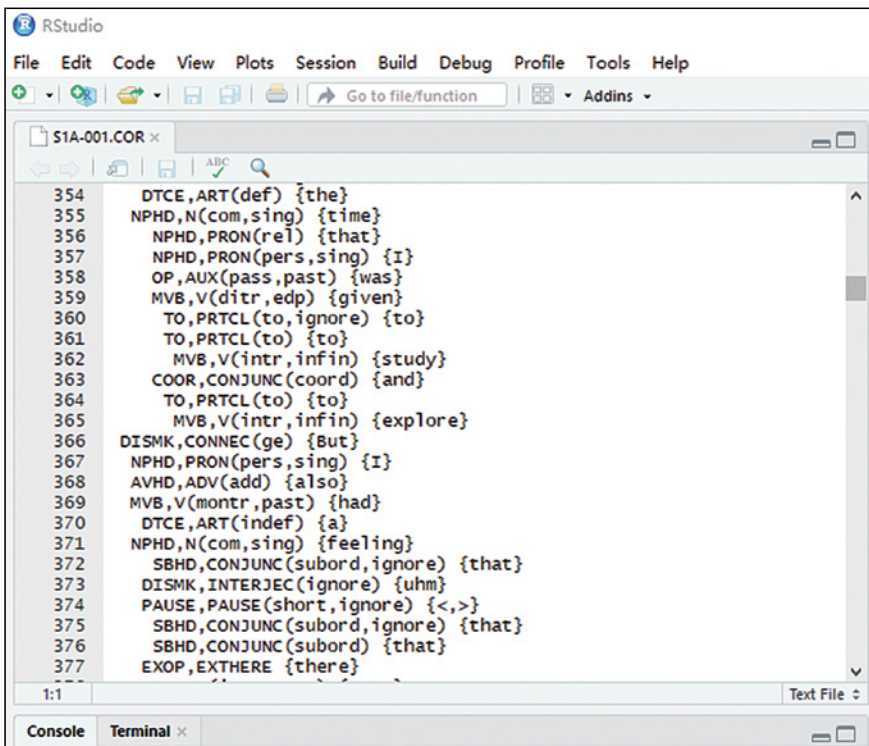
So if you look at the format that you have now, this corpus looks like this. This is the first file. In order to understand the build up here, basically there are two things you need to realize: One is that the words are always provided at the end of the line in curly brackets { }. What I did is I stripped out hundreds and thousands of lines of syntactic annotation, namely everything that is not also in a line that also contains a word.



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](https://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/m9.figshare.9611498>

But you can see here, for instance, here's the word *missing* [[in line 10]]. And here you see the syntactic annotation that comes with it. So *missing* in this sentence is a main verb, and it's a verb that's intransitive and used in the *-ing* form, and then here's the word [namely, MVB, V(intr,ingp) {missing}]. Here, the word is *what*. It's a noun phrase head, namely a pronoun, namely an interrogative pronoun [NPHD, PRON (inter) {what}]. Basically, what you have now here is, you don't have the parsed trees anymore, because like I said I didn't want to get into trouble by giving away a corpus that I shouldn't be giving away, but at least you have a part-of-speech tagged version of this. And the interesting thing for our application here is that at least some constructions can still be retrieved properly. And one of them is the one we're going to be looking at today, namely the ditransitive because, for instance, here you see an example of this.

Here in this line that is currently highlighted, line 359, the verb is *given*. That's the main verb. It's a verb that's used ditransitively. This is what we're going to use as a proxy for the construction, and then it's the past participle.



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
S1A-001.COR x
354 DTCE,ART(def) {the}
355 NPHD,N(com,sing) {time}
356 NPHD,PRON(reI) {that}
357 NPHD,PRON(pers,sing) {I}
358 OP,AUX(pass,past) {was}
359 MVB,V(ditr,edp) {given}
360 TO,PRTCL(to,ignore) {to}
361 TO,PRTCL(to) {to}
362 MVB,V(intr,infin) {study}
363 COOR,CONJUNC(coord) {and}
364 TO,PRTCL(to) {to}
365 MVB,V(intr,infin) {explore}
366 DISMK,CONNEC(ge) {But}
367 NPHD,PRON(pers,sing) {I}
368 AVHD,ADV(add) {also}
369 MVB,V(montr,past) {had}
370 DTCE,ART(indef) {a}
371 NPHD,N(com,sing) {feeling}
372 SBHD,CONJUNC(subord,ignore) {that}
373 DISMK,INTERJEC(ignore) {uhm}
374 PAUSE,PAUSE(short,ignore) {<,>}
375 SBHD,CONJUNC(subord,ignore) {that}
376 SBHD,CONJUNC(subord) {that}
377 EXOP,EXTHERE {there}
1:1 Text File
Console Terminal x

```

FIGURE 1



Obviously, a lot of constructions would not be retrievable just from these parse trees but ditransitive, since it is annotated like this, that is something that we can work with.

Strictly speaking, we could even distinguish active and passive ditransitives, as you can see here. We're just going to work with ditransitive in general to measure to what degree do certain verbs like to occur in that construction, or don't like to occur in the construction. So this is the format that we'll be working with. So that of course means a few things for extraction or for automatic extraction at least.

First, what we need to be able to do is we need to be able to extract the words. At some point, we read in that whole thing. We will want to ignore a lot of things, but we will want to recover that, ok, here's a word *was* and that it is given between curly brackets. One simple way, for instance, to get every word from the corpus would be to load all the lines and then delete everything from the beginning of the line until the opening curly bracket. And then in a second deletion, we would delete everything that is the closing curly bracket. With these two steps, applied to every one of these lines, we would be able to extract all the words. So that's one thing we will need to do, at some point, to find which verbs are used like that.

The second thing we need to do is we need to extract the ditransitives. What we'll be doing for this is we're going to look for a *V* because it means 'verb' followed by an opening parenthesis, followed by *ditr*. That will be our indication for, this is not just any verb, and it's a verb that is used transitively.

That's how we're going to make use of the annotation. I'm here again, using regular expressions of a certain type that help us identify this. Then we're going to look at two different kinds of association measures: One is using frequency and effect size and then we'll make a comparison to one other measure that does as well, and then we'll make a comparison to a third measure that only uses effect size, but not also frequency, the idea again being that for many applications that at least hope to be cognitively realistic, `[[it]]` might be more useful if we keep those dimensions separate.

It goes like this. How do we start with this? The first part in the script actually does something you haven't seen yet, but it is something that can sometimes be very, very helpful. So the first line here, it clears memory, so that's the same as always, but then you see a line here starting with `source`, and then a URL that leads to my website [<http://www.linguistics.ucsb.edu/faculty/stgries/exact.matches.2.r>]. So what the `source` function does is, the `source` function loads a file that could either be on your hard drive or, as you can see, on the web, and runs it in the background in R so whatever that file contains is then available in your R session. And so what this line does is that it loads a function

that I wrote, called `exact.matches.2` that comes with the second edition of my `[[corpus]]` textbook (Gries 2016). That's a function that makes it very easy to do concordancing in R. Doing concordancing in R in general is possible, but it's not really convenient, and that function does a lot of things that make it very easy for corpus linguists to work with that. Essentially, when you run that line, your computer will go to that website, download that script, run it in R and then after that this function is available for you as if you had installed it on your own computer.

One thing about that, though: Whenever you want to use that function, if you shut down R or R studio, you need to go there again. So it doesn't install it like a package that is then available always—it just loads it for one temporary session and as soon as you close R, typically at least, depending on your settings, it will be gone again. If you run that line, actually, nothing much is going to happen. If it was successful, you're just going to get a prompt, but it's not going to show you anything. But you can use that `exact.matches.2` function then to load things.

Let me just give you one brief example of what that function does: Imagine for a moment, you have a vector that contains two elements, like some random letter sequences, and then what the function does is you can look for something in a text. The output is then very useful for concordancing purposes. So the function is called `exact.matches.2`. It minimally takes two arguments: First, the string you're looking for. If you're interested in a word, then that would be the search word that you're interested in. If you're interested in some pieces of annotation, like `ditr` or whatever then that's what you will put in here. That second thing here would be a vector that contains the corpus data. In a moment, we will load each one of these five hundred corpus files and look for all the verbs in there that are used ditransitively and so on. So this will be the corpus file.

Then the nice thing that the function does is that something in R in general would not make that easy: It provides a list with a variety of types of output. Two of those outputs are of interest in particular for corpus linguists. The first one is the exact thing you were looking for. I'm looking for `bc`, and that's what it finds. That is interesting because in a lot of cases you might look for something that is not directly specified like here `bc bc`, but you might be looking for something like words ending in `ly` or something in English, because they are likely to be adverbs or something like that. Then you would see all those words here that were found, but nothing of the context. You only find exactly what you were looking for in this component.

The second part that is interesting is this part, the fourth output component. That one generates a concordance display. So it will take whatever it finds

and will show it together with its preceding and its subsequent context. So in exactly the way that you might be used to seeing output from concordancing software like AntConc, Wordsmith Tools or something like that. These two functionalities come in handy a lot if you do corpus work. I do all my corpus work with R and pretty much always with this function because either I want to see exactly what I was looking for or I want to see it in context so these two kinds of things are what are particularly useful for that kind of output. That's the reason why we're loading this function.

Now, the next line is again very familiar. We are defining for R the locations of all the corpus files. So we're saying `make_corpus.files` the content of the directory, namely, everything that's in the folder, `o3_data/ICEGB_sampled`. Like I said before, this is a corpus that contains five hundred files. So we're not just go to load fifteen files, whereas this time we're actually going to load five hundred files. Again, we're saying `full.names=TRUE`, so that all the file names, the whole path, the whole directory, is stored. If we run this in R, let me show you, what happens. Now we ran this, so we can see that we now have five hundred paths, for every single corpus file R now knows where it is. And these are the first six names of the corpus files to be loaded.

Since we have so many files, you can already imagine that we will, again, run a loop. So we will write code such that it can be applied to every single file collecting verbs and collecting ditransitives from every single file. That, of course, means we again need a data structure that stores the output from every file.

Last time around, what we did is we loaded a file, extracted all the words, and then we put them sort of in a collector structure for later. Then we loaded the next file, processed it, got all the words out and added it to the previous results, all the time. So this time, what we need is we need to find, for every file, all the verbs that are in there, and all the verbs that are used in ditransitives, because what we want to do is we want to create a data structure that allows us to construct a  $2 \times 2$  table for every single verb in the ditransitive. How often is a verb used in the ditransitive? How often is it used elsewhere? How often is this verb used in ditransitive? How often is it used elsewhere?

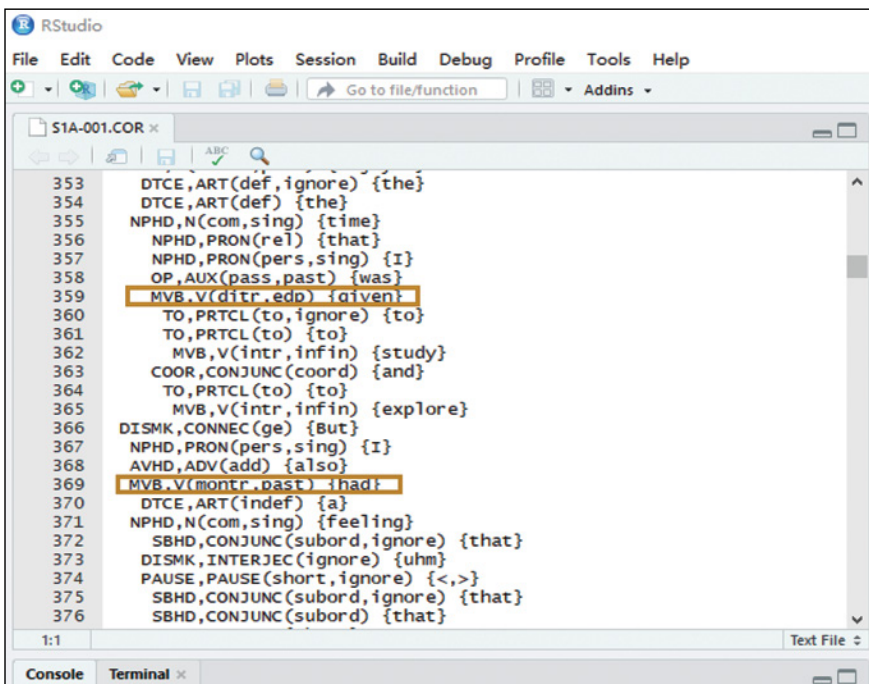
For the kind of association measure computations that I discussed yesterday, that's the kind of data we need. So the output is supposed to be, at some point, a vector that contains all the verbs in the corpus. It's probably going to be like 130,000 or something like that, 138,000. Then we also need a vector called `all.ditrans`. We're going to do it in the same way as we did this for the dispersion statistic. So what we are going to do is, we're going to collect all the verbs in the corpus in one vector, and then we're going to have a second vector that will say, for every verb, whether it is used ditransitively, yes or no. If the first verb in the corpus is *make* and it's not used ditransitively, then we'll have

*make*, not *ditransitive*. If the next verb is *tell* and it's used ditransitively, then we'll say this element will be *tell*, and this will be 'yes, used ditransitively'.

So if there are 138,000 verbs in the corpus, then this thing will contain 138,000 verbs. And this thing will contain 138,000 *yess* or *nos*, depending on whether the verb is used transitively, yes or no. Then, we can generate a table that says for every verb how often is it used ditransitively and how often it is not. Just like last time we did, we created a table that said, for every word in the corpus per file, whether it is the word in question, like *enormous* or *staining*, or not.

That's essentially what we're going to do. Now much of this is actually the same as before. So I'm actually gonna run it in R and then discuss it in a file.

So, pretty quick. What are we doing? We are again beginning with the loop, this part is all the same, so `for (counter in 1:500)`, because we have 500 files. When R gets here the first time, `counter` is set to 1. And then `current.corpus.file` is the result of loading `corpus.files[counter]`, right now the first (file). And that is the character file that contains character strings separated by line breaks, and don't give me any output. So after this part, the first file has been loaded. Now we need to get the verbs.



```

353 DTCE,ART(def,ignore) {the}
354 DTCE,ART(def) {the}
355 NPHD,N(com,sing) {time}
356 NPHD,PRON(rel) {that}
357 NPHD,PRON(pers,sing) {I}
358 OP,AUX(pass,past) {was}
359 MVB,V(ditr,edp) {given}
360 TO,PTCL(to,ignore) {to}
361 TO,PTCL(to) {to}
362 MVB,V(intr,infin) {study}
363 COOR,CONJUNC(coord) {and}
364 TO,PTCL(to) {to}
365 MVB,V(intr,infin) {explore}
366 DISMK,CONNEC(ge) {But}
367 NPHD,PRON(pers,sing) {I}
368 AVHD,ADV(add) {also}
369 MVB,V(montr,past) {had}
370 DTCE,ART(indef) {a}
371 NPHD,N(com,sing) {feeling}
372 SBHD,CONJUNC(subord,ignore) {that}
373 DISMK,INTERJEC(ignore) {uhm}
374 PAUSE,PAUSE(short,ignore) {<,>}
375 SBHD,CONJUNC(subord,ignore) {that}
376 SBHD,CONJUNC(subord) {that}

```

FIGURE 2

Let me remind you of what the format of the file is like. It looks like this [referring to Figure 2]. Like I said, we will operationalize ‘verb’ using the annotation that’s in the corpus. Namely, we’ll work around this V here, because that means verb, but only V if it’s in front of a parenthesis because we want to make sure that there’s not any other V in the annotation, like this V here [refers to Figure 2, line 369] for main verb, or if there’s an adverb (ADV), you know, we don’t want that one, so we’ll try to make sure that we only capture those cases that are relevant for our search.

This is what I’m doing here. I’m creating an object called `curr.verbs`, so that will be all the verbs in this file, which is the result of searching for something—I’ll explain this in a moment—in the `current.corpus.file`, `perl=TRUE`, use Perl-compatible regular expression, so this contains some regular expressions. Then, `value=TRUE`, return what you find. The thing about `grep`, this finding function is, if you don’t say `value=TRUE`, it will only tell you where in the file it is, but it will not return what it found. So without `value=TRUE`, `grep` will tell you, I found something in the second line, but it’s not going to tell you what it found. If you say `value=TRUE`, it will tell you what it found. And we here of course we want the verb, so that is what we are going with.

Now, what does this expression `[\\bv\\(]` mean? The simplest thing in the middle is the `v`, that’s the verb tag. And then we have two backslashes and the opening parenthesis. That needs to be explained a little bit: A parenthesis is a character string that also has a regular expression meaning. What a parenthesis does, it defines a unit for a regular expression. But what we want R to do here is we want to actually really use a parenthesis: We don’t want R to interpret this as a regular expression—we really want to find a parenthesis. So the way to do that in R and in many other programming languages is to prefix, in R’s case, two backslashes. That means this opening parenthesis is not a regex, I really mean that character. Another way which might be useful: remember yesterday, we used periods (.) to say ‘anything’, at the beginning of the line. But of course there might be a situation where you actually want to find the period. What you would need to do is you would need to tell R “don’t use the period as meaning ‘anything’, use it as meaning ‘a period’”. The way to do that would be the same as this: You would put two backslashes in front of the period so that R knows that the user doesn’t mean ‘anything’, it really just means ‘period’. So that’s what this does. So we’re looking for a V in front of an opening parenthesis. And this part here, `\\b`, that means that before that V, there can’t be any other letter or number. There’s actually something else that it also means, but it doesn’t matter right now. But why am I saying that? I’m saying that so that it cannot find ADV for adverb, because in ADV there *is* another letter in front of

```

353 DTCE,ART(def,ignore) {the}
354 DTCE,ART(def) {the}
355 NPHD,N(com,sing) {time}
356 NPBD,PRON(reI) {that}
357 NPHD,PRON(pers,sing) {I}
358 OP_AUX(pass,past) {was}
359 MVB,V(ditr,edp) {given}
360 TO,PRICL(to,ignore) {to}
361 TO,PRICL(to) {to}
362 MVB,V(intr,infin) {study}
363 COOR,CONJUNC(coord) {and}
364 TO,PRICL(to) {to}
365 MVB,V(intr,infin) {explore}
366 DISMK,CONNEC(ge) {But}
367 NPHD,PRON(pers,sing) {I}
368 AVHD,ADV(add) {also}
369 MVB,V(montr,past) {had}
370 DTCE,ART(indef) {a}
371 NPHD,N(com,sing) {feeling}
372 SBHD,CONJUNC(subord,ignore) {that}
373 DISMK,INTERJEC(ignore) {uhm}
374 PAUSE,PAUSE(short,ignore) {<,>}
375 SBHD,CONJUNC(subord,ignore) {that}
376 SBHD,CONJUNC(subord) {that}

```

FIGURE 3

the V. But in the tags that we're interested in, it's ,V so the comma is not a letter. So this is my way of ensuring that the V is not the last character of a multi-letter tag, like ADV for adverb, but there's a comma in front of it or a space in front of it, so that's really a verb tag. Technically, the `\\b` means word boundary so the V is not part of the word, because there are letters in front of it—it's the beginning and the end of a tag at the same time.

So that makes sure that this V [referring to `v(ditr, edp)` in line 359] gets found because there's no letter in front of it. But for instance, if we only look for the V, this one [referring to `MVB` in line 359] would not be found, because there is a letter in front of it. That's how that works. We don't have time to discuss regexes in great detail, but at least it gives you an idea of how this would be used.

Now, after we've run this part, we have all the verbs in that corpus in this thing. Actually we have them in there with the annotation. What `grep` returns when it finds something is the whole line. That means: this whole thing is what's going to be in `curr.verbs`, so not just the word per se, but also the

annotation. That's actually great, because what it means is, if we now want to find the verbs that are used ditransitively, we can take this as input and check whether it contains the ditransitive tag. We didn't just look for a verb at the end of something, no, we look for that whole long string, and that means that we can now look in that for whether it contains the ditransitive tag. That's what the following line does.

We have created `curr.verbs` here, now we're creating `curr.ditrans`. What are we looking for? We're looking for whether `curr.verbs`, the thing we just found, contains `v` followed by `di tr`. So we have the whole corpus file, we find the verbs in there, and then in those verbs, we look, is there also a ditransitive tag, yes or no? That also of course makes things much faster. In this search, where we look for the whole verb and everything, this goes through the whole corpus file, so potentially, hundreds of thousands of lines. This one only goes through the verbs. So we're looking at, maybe just ten percent of the original size, so that makes it much faster.

But now, there's one little trick that I am employing here already. Here we were looking for verbs, and the function was `grep`. Here I'm using `grep1`. What `grep1` does is, it does not return where it finds something—it returns `TRUE`s and `FALSE`s whether it finds what you're looking for or not. So what that means is, let's say, `curr.verbs` has only five verbs in it, because I have five fingers. If you then run this line, then `grep1` will return for every one of the five verbs `TRUE` or `FALSE`, depending on whether it's ditransitive or not. If let's say only the fourth verb is used ditransitively, the `grep1` search will say, `FALSE FALSE FALSE TRUE FALSE`.

So this is our way how we create a vector with all the verbs and a vector with all `yess` and `nos` that says whether that verb is used ditransitively or not. Because then at the end, we are doing what we've always done so far: We say `all.verbs` is the combination of whatever `all.verbs` was so far plus the stuff from the current search, and `all.ditrans` is whatever it was before plus the results from the current search. Once we're done with this loop, we have again like approximately one 138,000 or something verb forms, and 138,000 `yess` and `nos` that say whether something is used ditransitively or not.

Let's actually check this out. I already ran it here in the background. `length(all.verbs) = 140,880` and then the same thing with `length(all.ditrans) = 140,880`, so those are equally long.

Let me show you just 15: These are the first 15 verbs in the corpus using the annotation. So this is annoying, there's not a single ditransitive in there. But you can see complex transitive, intransitive, monotransitive, all the syntactic annotation in there. So these are the first 15, and there's no ditransitive in

there. So I hope that the next, if I now show you the first 15 elements of all ditransitives, they should say `FALSE`, because none of those is used ditransitively. So let's see, this is R telling you none of these first 15 is in fact ditransitive [15 `FALSEs`].

Maybe a quick question, how do we find out how many ditransitives are there in the whole corpus? There's two ways. One is this. So there are 1,841 ditransitives in the corpus and 139,039 other verb uses. We just tabulate how many *yess* and how many *nos* essentially are in that vector. The other possibility you might have is the one that remember I told you the other day, `TRUE` and `FALSE` are 1 and 0 so you can just sum it up.

So what do we want to do next? As you've seen, the output showing the verbs is actually not that great, I mean, not in terms of how do we process this further. There's a lot of crap in there that we actually don't need. Actually, we don't need anything other than the verb here. Since this is not ditransitive, we don't need the annotation. The same here for any of these verbs, we actually really only need the verb form now, because we want to create a table that says for every verb form, is it used ditransitively, no or yes? There's a ton of spaces here, all sorts of other things we don't want. So now the next thing we want to do is, extract the word right out of these two curly brackets. That uses a regex that I won't have time to explain here very well.

I'm just going to walk you through it by highlighting it to the extent that I can. So what this regular expression `(?<={ } [^])+` does is it looks for this. What does the `+` mean? We talked about that yesterday, 'one or more'. So the question mark would mean 'zero or one' and a `+` means 'one or more'. And you can see the square brackets again. The square brackets were a character class. And remember that within a character class, this caret sign `^` means not. So this means, try to follow along to the extent that you can see it, one or more cases of something that's not a closing curly bracket.

That's what the verbs are [referring to Figure 4, line 359], they are one or more things that are not a closing curly bracket. Another verb here *explore* [referring to Figure 4, line 365]. The verb is one or more characters that's not a closing curly bracket. But what do all the words have in common? They have in common that they are coming after a curly bracket, an opening one: Every single word here is preceded by an opening one and ends with a closing one.

So what the regex says—let's try to follow along the gesturing again: one or more things that are not a closing curly bracket, *if*, when you look to the left, you see the opening curly bracket. It's like this: One or more things that are not a closing curly bracket [referring to Figure 4, line 359], but *only if* when you



```

353 DTCE,ART(def,ignore) {the}
354 DTCE,ART(def) {the}
355 NPHD,N(com,sing) {time}
356 NPHD,PRON(reI) {that}
357 NPHD,PRON(pers,sing) {I}
358 OP,AUX(pass,past) {was}
359 MVB,V(ditr,edp) {given}
360 TO,PRTCL(to,ignore) {to}
361 TO,PRTCL(to) {to}
362 MVB,V(intr,infin) {study}
363 COOR,CONJUNC(coord) {and}
364 TO,PRTCL(to) {to}
365 MVB,V(intr,infin) {explore}
366 DISMK,CONNEC(ge) {But}
367 NPHD,PRON(pers,sing) {I}
368 AVHD,ADV(add) {also}
369 MVB,V(montr,past) {had}
370 DTCE,ART(indef) {a}
371 NPHD,N(com,sing) {feeling}
372 SBHD,CONJUNC(subord,ignore) {that}
373 DISMK,INTERJEC(ignore) {uhm}
374 PAUSE,PAUSE(short,ignore) {<,>}
375 SBHD,CONJUNC(subord,ignore) {that}
376 SBHD,CONJUNC(subord) {that}

```

FIGURE 4

look at *given* and you go one step to the left, you'll see the opening one. That's how we're now picking the word out of these two curly brackets.

So let me show you that in R: `all.verbs` right now is this. Now we're picking out the verbs. So everything from the annotation is gone because we said 'only give me this, namely stuff that's not the closing curly bracket and is to the left of an opening one' ... so now, that looks nice. It looks like something we can actually use without too much annoyance.

Then the only other thing we want to do, I'm taking shortcuts here, but I want to show you one example. There are sometimes cases where a word is interrupted by some annotation, and this annotation I'm taking out. I don't see a good example here right now; let me see whether there is a good example in another corpus file. Anyway, sometimes you will find cases. This is what the next line does in the code.

This is how well you need to know your corpora.

Sometimes you will find words that have this annotation [`<1>`] in them, an opening angular bracket, an `1`, and a closing angular bracket. Obviously, we want to get rid of those. So what we're doing here is now that we've extracted all the verbs, I'm taking out any of these annotation cases and replace them by nothing, just to make sure that the annotation doesn't mess up our counts. If

there's *given* or something, but some of these cases have < after the *i*, then obviously, R will think it's a different verb so I'm cleaning up the corpus like that to make sure this doesn't interfere with our counts.

Now it's actually really simple, because now we do what we did yesterday with the dispersion script and that is, we create a table that is called `verbs.by.ditr`. So we're cross-tabulating every single verb in the corpus, all 140,000 with whether it is used ditransitively, yes or no.

And if we do that, the upper part of that table looks like this, which is not great. That is because this table is sorted alphabetically, so these punctuation marks, the apostrophes, come first. But so you can see there are three cases of *'d* probably for *would* and three of them are not used ditransitively, zero ones are used ditransitively. There's one case like this *I'd rather* and it is not used ditransitively. But this looks like very useless, because we have tons and tons of verb tokens what we don't even get to see the ditransitives here. So what we're going to do now first is reorder the table such that ditransitives go at the top. So this is what this line does.

So, create a new version of `verbs.by.ditr`, which is `verbs.by.ditr` the old version, but now order it by the column called `TRUE` in decreasing order. That means in a moment when we do this again, we will see the most frequent verb used ditransitively then the second, then the third, and so on.

Let me show you, and there we go: *give* is used 206 times not in the ditransitive and 237 times in the ditransitive. *told* is used 103 times not in a ditransitive, 219 times in a ditransitive and so on. Now we can easily see for every verb that is in the ditransitive at all how often it is used there. For instance, look at *ask*, it is in the top 10, but actually *ask* is way more often used outside of the ditransitive than in the ditransitive. Whereas all the top three, at least, are used more often ditransitively than not. The first switch is here. So *given* is used more often not in a ditransitives than in ditransitives. But if you just use frequency, then this would actually be your main result: The frequency of every verb in the ditransitive, you could make that into a percentage or something, but that would be it. But of course we want to do association so we'll have to go a little bit beyond that.

Now, the simplest way to compute an association measure on this is actually one that is hardly ever used by people. I'm not a hundred percent sure why, because I will show you that the results are pretty much identical to what everyone is doing. I will give you this very simple computation way, because I think it's actually much better than what a lot of people are using.

What I'm doing here is I'm creating a vector `assoc.strengths`. What I do is I run a `chisq.test` on the table that you just saw. I'm doing a chi-squared test on that table. Now, what does that do? Let me show you what that does.

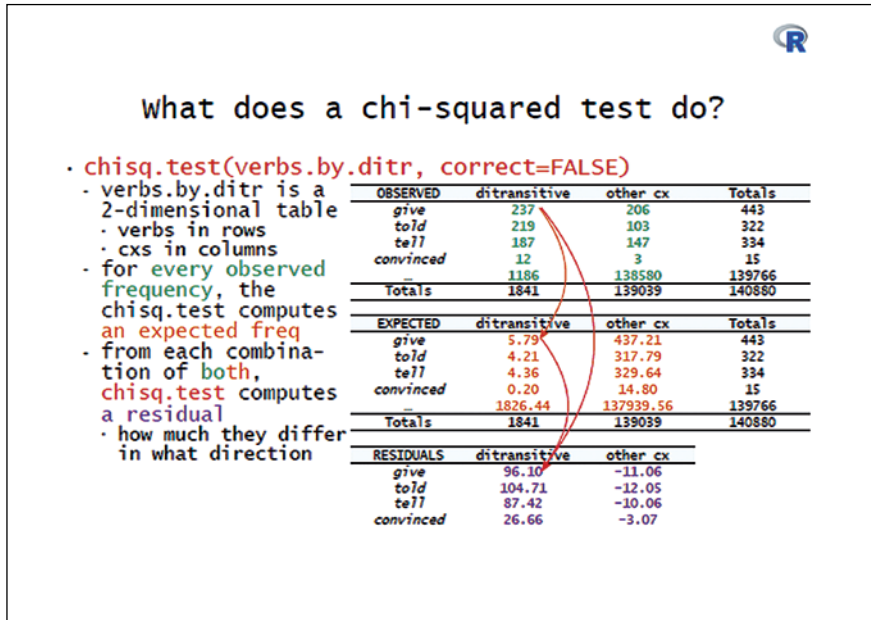


FIGURE 5

This is the R's code that we're running, `chisq.test`, on this table, this table of all the verbs and ditransitives, no and yes, and here I'm suppressing a certain continuity correction, doesn't matter right now. This is the input table for, this is what you saw right now. The ditransitives were in the second column for you. *give* is used ditransitively this often [[237]], in other constructions that often [[206]], which means this is the overall total of *give* [[443]]. *told* is used 219 times in the ditransitive, 103 outside of it, so that's the total. And then this is the overall frequency of the ditransitive. In this table, I'm showing you the first four verbs *give*, *told*, *tell*, *convinced*, and then this [...] is like every other one, and then summing up to that total. That's the input to this `chisq.test`. So then what the `chisq.test` does is it computes for every observed frequency, for every green number, it computes an expected frequency.

So for every observed frequency, it computes an expected frequency. What the expected frequency says is the following: If there is no relationship between the ditransitive construction and all the verbs in it, how often should *give* be in there, if there wasn't this super strong semantic compatibility of the meaning of the verb *give* and the meaning that we attribute to the construction? As you can see, if the verbs were randomly distributed across constructions, there should be approximately six uses of *give* in the ditransitive, but there's more than 200 there. This number here is forty times as large as expected.

The same with *told*. If there was no relationship semantic or otherwise between the verb form *told* and the constructions, there should be four instances of *told* in the construction. But there's approximately 50 times as many attested. The way the `chisq.test` computes the expected frequency is as follows. You just kind of need to follow along. So it takes the row total, multiplies it by the column total, and divides by this [refers to 139,766]. So this number here, 5.79, you get that, if you multiply 443 times 1841 divided by that [refers to 139,766]. And this number here, 4.36, is for *tell*. So you take the row total of *tell* times the column total here, and divide by that, then you get that number [refers to 4.36]. That's how the chi-squared test says sort of 'this is the expected frequency'. And the way this works is now, if you look at these numbers, so what that does is: the ratio of this number to that number is the same as this to this, is the same as this to this, is the same as the total to this total. So it basically means 'for every verb, the frequency of ditransitive versus other is the same'. That's what the chi-squared test does. Then from the expected and the observed frequency, the chi-squared test computes a so-called residual, which is this number [[96.10]]. It is now doesn't matter how that one is computed, it's not complicated, really, but [[that goes beyond]] what I want to talk about today. And the residual says two things: First, it tells you whether the observed frequency is greater than the expected one and it tells you that by means of the sign. So when this residual is positive, then the observed number is higher than the expected. 237 is higher than 5.79, so this 96.10 is positive. But look here, 206 is less than expected, so this one, -11.06 is negative. So, all of these, i.e. *give*, *told*, *tell*, *convinced*, are preferred in the ditransitive.

The second thing it tells you that the more the residual deviates from zero, the stronger the effect. This is already interesting because you can see that in our analysis here right now, the verb form *told* is actually more attracted to the ditransitive than the verb form *give*. That in part has to do with the fact that we didn't lemmatize so we're not treating all the forms of *give* together, we're not treating all the forms of *tell* together. Otherwise, the result would be somewhat different. But so now, this you can use to rank-order the verbs in terms of their attraction to the construction.

So if we do this in R now, I computed this, then I sort this: These are now the top thirty collexemes of [the ditransitive].

As you can see, they all make sense in a very informal way. I mean they're all compatible with transfer or with communication. I mean, *told*, *give*, *tell*, *given*, *gave*, *telling*, *gives*, *tells*, *giving* ... all the same. Then things like *convince*, *offer* is the offer to transfer, *sent* is 'transfer across a distance', *assured* is a communication verb, *show* can maybe be seen as metaphorical transfer, *ask*, *sent*, *cost*, all these items that all these analyses of the ditransitive have always prioritized do

in fact come out here as the top collexemes. This could be the range of numbers that you report if you do a collexemes analysis using these Pearson residuals as an association measure. Again, most people don't do that, but it's actually simpler to compute than what nearly everyone is doing and the message or the conclusions are pretty much exactly the same.

Now, two small things. One is that I do want to prove that point that this test that you will not see in use by anyone, actually does return the same results as what everyone is doing, namely using the log-likelihood coefficient. So the code that follows computes log-likelihood statistics that are very similar to what people use when they use Fisher-Yates exact test as a script. And this one actually takes a while to run, it could be two or four minutes because it has to compute a whole bunch logistic regressions. But what I want to show you is the plot that shows how the measure that I've showed you and the one that everyone is using and how they are related.

On the *x*-axis, you have the test that we computed, on the *y*-axis, you have the test that most people are using. As you can see, in general, there's an extremely high correlation between the two. It's very easy to draw just with your finger, you see, there's a line that captures everything pretty well.

In fact, I show you in the code how high that correlation is. So the correlation between what everyone else is doing and this simplified version is nearly perfect. If you round nicely, it is 0.99. But it is, actually, probably, I think a better way than what most people are measuring.

The other thing I wanted to show you very briefly is the log odds ratio computation. The reason for that is that the measure that I showed you does what we actually do not usually want to do, at least not if you follow what I said yesterday, namely compute an association measure that conflates frequency and effect size.

The log-likelihood ratio that most people are using, it does reflect frequency quite strongly like what we saw in the talk yesterday. And the measure I'm using here, the chi-squared test does as well. So what if you do not want that, what if you actually want to separate the two? Let me very briefly refresh your memory on how the log odds ratio is computed.

It is computed by taking the frequency of the verb when the construction is there, divided by the frequency when the construction is there but the verb is not. And you divide that by the same for when the construction is not there. I mean you can actually just visually memorize this, you know: if you have a 2×2 table, with these four cells, you just divide this  $\frac{80}{607}$  by this  $\frac{19}{137958}$  is 0.1318, and this  $\frac{19}{137958}$  divide by this  $\frac{80}{607}$  is that  $\frac{19}{137958} \div \frac{80}{607}$ , and you go and make that division [and get 6.8638]. Yesterday, we saw in the results that this

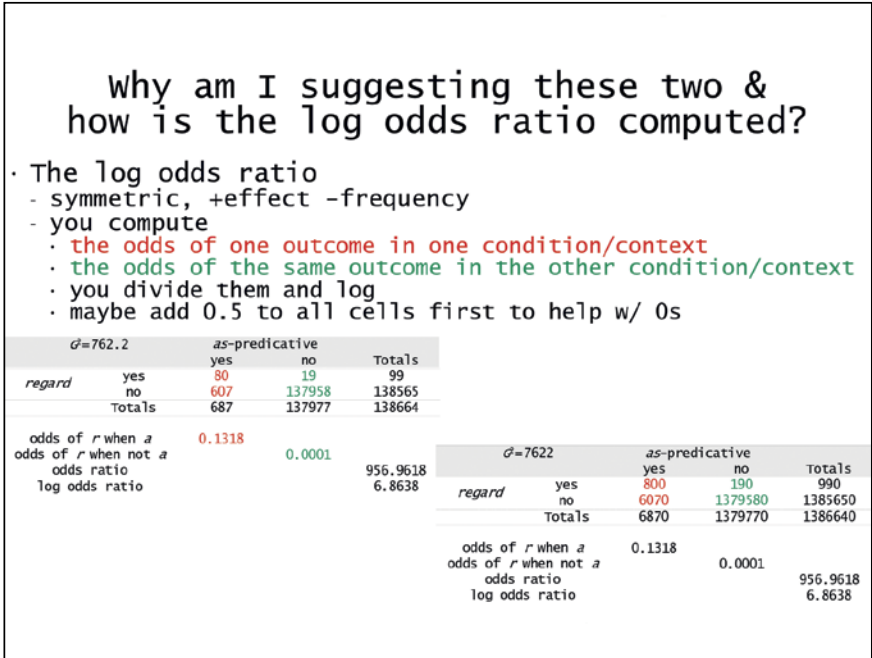


FIGURE 6

measure does not react to frequency because as you can see here, the log odds ratio is 6.8638—if the corpus is ten times as big, it's still 6.8638. So it would be nice if we can do something like that also in our R script, and of course we can.

So that's the remaining part that the code here does. So this, too, will actually take a moment because I am computing it in a way that is theoretically different, but while it's computing in the background, let me show you the results.

This is quite interesting, but actually it also shows that what we are doing is not stupid. These are the verbs that are sorted according to how strongly they are attracted to the ditransitive without the effect of frequency. So *accorded* is the strongest, and then *teaches*, and then *convincing*, and then *overpaid*, none of those are what we would expect. We would expect *give*, *tell*, *show*, *offer*, *ask*, I mean, those are nowhere to be seen until we get to here [*told*, *tells*]. So that tells you that all the high results that we usually see, for *given*, *tell* and so on, those are in part driven by frequency. Because if you only go by association, they actually don't score that highly. Other verbs like *accorded* are actually more strongly attracted to the ditransitive if we take frequency out of the equation.

So if we visualize this, that becomes even clearer. And you've seen a version of this plot yesterday. It looks like this. As before, what we have on the x-axis

is frequency. What we have on the  $y$ -axis is association. So here's things like *convincing*, really not very frequent at all, but probably used in the ditransitive all the time, exclusively, just not often enough.

Then on the other side, you have things like *get*, very, very frequently used in the ditransitive, probably something like, you know, *I got them something as a birthday present* or whatever. But actually, the attraction is relatively low. And that's because *get* is an overall very frequent verb, it shows up everywhere. So the fact that it's frequent in the ditransitive doesn't mean much if it's not also strongly attracted to it, which it is not. The value is actually really close to zero. In the plot like this, what you always want to look at is the verbs at the top and on the right. You want to look at this area, because this is where words are that are relatively frequent, but also relatively high up. So it's always this area. So ideally, you would be looking at this here, super frequent *and* super strongly attracted but of course, that is never going to happen so basically, you go from the top right corner and see what are the things that I'm hitting first.

Then of course we do see the usual suspects, *told, give, tell, given, gave, telling, tells, assured, convinced, offered, cost*, all these kinds of things. Again, you can see that the correlation between frequency and association is really not that high. I mean, yes, there is a downward trend like this. But it's not nearly as strong as what the plot that I showed you before where we had this super straight line. Again, I hope this makes clear why being able to separate these two pieces of information is potentially and theoretically very useful, because for our minds frequency is not the same as association.

As we know, if something is very salient than a single repetition or a single occurrence of something can lead to learning already. You hear a new word and you immediately know it. Even very young children can do that, even when they're not taught words explicitly. A single salient mention of a word in a context can lead to retention for one or two weeks even without repetition. That's because the association was perfect, and the salience was very high so it doesn't matter if frequency is low. At the same time, of course, we can learn things even that are not particularly salient if they occur often enough together. So even just cognitively speaking, it's not obvious at all that these two things should be always related. So measuring association in a way that is separate from frequency is something that will allow us to get a clearer picture of what's going on.

Obviously, discussing this in R always is a little bit difficult given the short amounts of time we have, but I hope you have at least seen the overall structure of the script even if we didn't have time to go into much detail about regular expressions. Let me know if you have any questions. Thanks.

## On Context

All right, thank you very much. The title has actually changed a little bit because of some last-minute prep, but I will still talk about quantitative stuff and I will still talk about corpus methods. The topic of this talk basically follows up on the preceding ones where we basically built up this idea from not just doing frequency, but also recency, and then also association. And today, in this talk, I want to talk a little bit more about the role of context and the way in which corpus-linguistic approaches can be applied to context to quantify certain things that are relevant to learning, processing, acquisition and so on.

So as we've seen earlier, we started by talking about frequencies and I made this connection that every cognitive and usage-based linguist has been making, namely that there is a relationship between token frequencies on the one

**Brief recap**

- we have talked a lot about frequencies - type and token - earlier this week; on the whole
  - token frequency was related to entrenchment
  - type frequency was related to productivity
- however, token frequency was found to be largely epiphenomenal (Baayen 2010)
- instead, a bunch of other factors - local syntactic & morphological cues - was found to be more important for matters involving lexical access
  - we already dealt with recency (priming & dispersion)
  - we already dealt with contingency
- now, we're dealing with
  - both a narrower and a broader view of contexts
  - type-token distributions
- but also with the often non-discussed limits of what context is/does

FIGURE 1



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611552>



hand and entrenchment on the other. At the same time, we talked about how type frequency was related to productivity and so on. But then I discussed a few times the kind of results that were offered by Harald Baayen's (2010) huge analysis that showed the token frequency was largely epiphenomenal. And in his study, I pointed to that one time, he actually found a whole bunch of other factors to be way more important than frequency. They were correlated with it, but they were also more important. And, at that time, I just said "local syntactic and morphological cues", so things that happen syntactically and morphologically around the words and questions. And so we looked at something like recency so far; we looked at contingency, and today we're going to take two kinds of views on context: one sort of a very narrow view and one much broader view of context. And at the second—at the same time, we're going to have a brief look at type-token distributions again, and revisit this notion of entropy and what it can offer us.

In a second, a bigger part, at the end of this talk, I want to discuss a particular case study that, at the same time, wants to damp down a little bit the effect context might have. As you will see, I will provide some quotes that show how usage-based linguistics has been talking about the role of context, how it has been talking about all the different dimensions that might go into exemplar-based models of linguistics. And I want to show a very simple set of experiments by a PhD student of mine that were designed to test how far can we actually push that kind of agenda, that kind of idea that basically we pick up

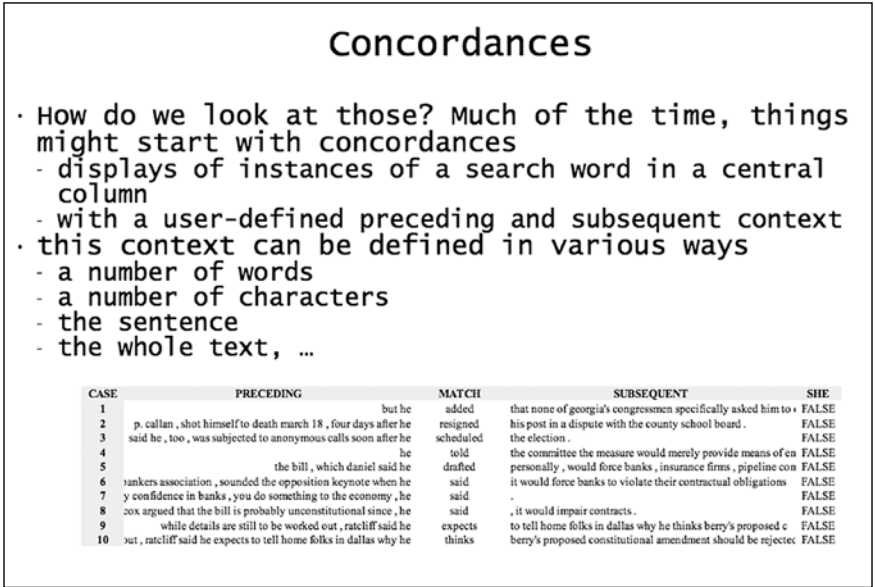


FIGURE 2

on everything in context of a linguistic experience and put that exemplar into a certain position in multidimensional space and, to already anticipate the results, we will see that context, the way it is at least talked about in usage-based linguistics, is not as powerful and we need to be more careful in terms of distinguishing or motivating which dimensions play a role.

Now, the simplest way to look at a corpus linguistic context is using a concordance, of course. How do we look at those? Most of the time, we look at concordance displays where we have a search word or search construction or something like that in a central column and then we have a user-defined preceding and subsequent context. Most of you probably have seen that in some way, if only in an online concordancer. And of course, we can define context in a variety of ways. So in the practical sessions that we’ve had over the course of this week, a lot of times it was just the line in the corpus file and a lot of times that of course conveniently corresponds to a sentence. But of course, it can be all sorts of other things as well: You can use a number of words, a number of characters, sentences and the whole text—theoretically, all of those are possible. All of those typically lead to a display of some sort of type like this, so where we have matches, something we were interested in, and then the preceding context and the subsequent context, typically a case number, and then we can read the concordance for whatever kinds of this information that we want. This case here, you see the concordance is really a very raw one, because you can see in those lines here that actually the tags are still in there, every

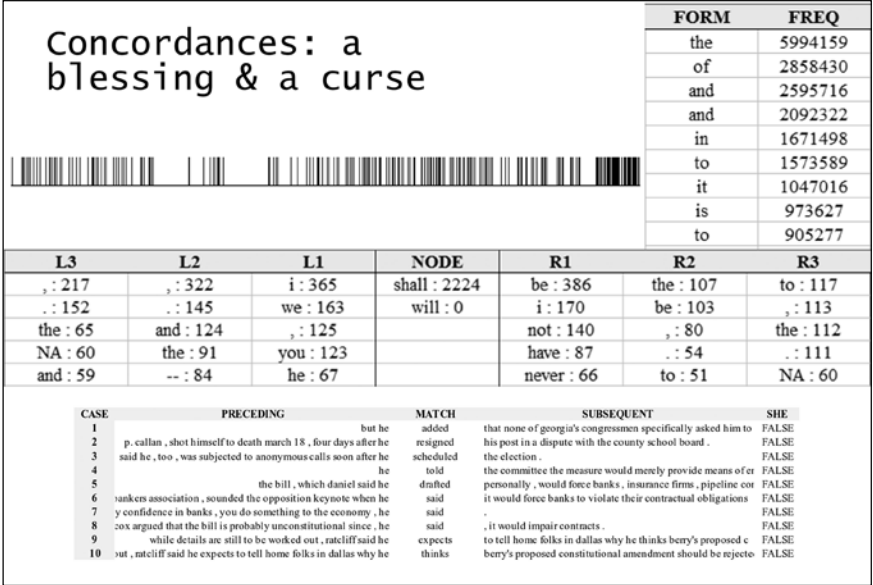


FIGURE 3

## Concordances: a blessing & a curse

- Concordance displays provide the most context of all corpus-ling. methods
  - you don't just have frequencies of decontextualized words
  - you don't just have dispersions of decontextualized words
  - you don't just have (tables of) collocates of words in slots around a word
- no, you see all of the context (limited by your data and their annotation, not so much by you by choice)
- that's both the main strength and the main weaknesses of concordancing: you see everything, all the context ...
  - great for detail and fine-grained annotation
  - not so great for automatic post-processing

FIGURE 4

word here is followed by a tag, just like we've seen in multiple practice sessions so far. Of course, those can be edited to take the tags out of the context to make things more readable and so on. But still, the idea would be that we have some sort of tabular displays like that we then annotate.

Now, given everything we've seen so far, concordance displays provide the largest amount of context. We don't just have frequencies like this which are completely decontextualized. We don't just have dispersion, which is a little bit more contextualized but not much. I mean it's contextualized in the sense is that this word occurs here in a certain context, but of course we don't know what the context is. We also don't have just collocate counts, or something how frequently does something happen but we actually see the whole thing.

And that of course is both a strength and a weakness in a sense, because it's really, really great for fine-grained analysis, for detailed looks at concordances and the context and so on. But of course, it also means it's really not so great for anything having to do with automatic post-processing, simply because the volatility, or the versatility, of language use will make it very difficult to find patterns in there that we can reliably annotate automatically as many of you will know. So how do we deal with this?

Well, a lot of times what we can do is we need manual post-processing. A lot of things defy any kind of a standard automatic annotation. So what you might need to do is, for instance, you need to search for something automatically, using a concordancing tool to get maximum recall, to make sure that you find everything that might be a hit. But then you need to go through all those hits manually to discard everything that fits the structural description that you

### Common practice in concordancing, to get precision & recall right

Given that annotation is not always available or precise enough, a lot of manual post-processing might be required: you may need to

- search for something automatically to get maximum recall
- discard false hits manually to get maximum precision

often compromises might be necessary ...

the dative alternation in non-annotated learner data

- you might get a list of the 20 most freq verbs in the alternation in annotated native-speaker data

- search for all those verb forms in learner data
- discard all those hits that are not the dative alternation

prenominal adj. order in non-annotated learner data

- you might get a list of all adjectives out of the BNC
- tag all occurrences of adjectives in non-annotated learner data
- look for sequences of two things tagged as adjectives
- discard all hits that are not two adjectives in front of a noun

FIGURE 5

were looking for, but is not actually a hit. For instance, in one case, a colleague and I were looking at the *into*-causative construction in English, *to trick someone into doing this, to force someone into doing that, to bully someone into admitting that* or something like that. So always this: a verb, direct object, the word *into*, and then an *ing* form. And we used the newspaper corpus that we had at the time because it was the biggest one we had, which was not tagged. So we were not able to look for *into* plus a verb in the gerund.

So what we looked for is *into* following anything by *ing*, anything ending in *ing*, because we didn't have any linguistic annotation. So we had 16,000 hits, and then we have to read them all to decide which of them were *into*-causative and which were not. So that is the life of a corpus linguist a lot of times. So we still actually got more than 9,000 out of it, [which] was a pretty good rate. But the idea is [that] a lot of times you need to do a general search first, and then weed things down to what is actually relevant. And it also means that sometimes you will have to make compromises that "compromise", no pun intended, the recall. So, for instance, someone else and I did a study on the dative alternation in learner corpus data that were not annotated. So we were not able to look for any occurrences of parts of speech or parts of speech in word combinations or something like that. So what we then ended up doing is the following: we took native speaker data that were nicely annotated and then we picked the twenty most frequent verbs in that alternation. Then, we looked for those verb forms in the learner data. So obviously that means we had to

look for *give*, *tell*, and *show*. And so we got all the verb forms from those in the learner data. Then again, we had to read them all manually to decide is this an example of the dative alternation or not. So what we ended up with then was a list that had maximum precision—all the hits were manually read and checked for whether they are a valid example or not—but of course, we did not have maximum recall because any verb that was not among the top twenty, but still can be used in the dative alternation, that, we did not find. So ideally, then you would hopefully have a lot of coverage from the twenty or thirty most frequent items, but you won't get maximum recall.

The same thing here. This would be another kind of compromise. That's also not going to find everything, but hopefully most of it and very precisely.

So in another case, we looked at prenominal adjective order in learner data. So the question was, how do you decide which of two adjectives in front of a noun goes first. Do you say *the old brown car* or do you say *the brown old car*—which of the two? Our learner data were not part-of-speech-tagged. We couldn't look for adjective-adjective-noun [combination]s. So we simply extracted all adjectives in the British National Corpus and then we looked through the learner data and tagged every word that looked like one of those from this list as an adjective. So we basically tagged our learner corpus but just for adjectives. And then we looked for whenever there's a sequence of two, and then read them manually.

So again, if the learners used an adjective that was not on the BNC, we will not have found it so we do not have perfect recall: We're not finding everything, but at least we're going to find probably 95% of them and it's maximum precision because we read every single item. Again, a lot of times, of course, [it is] boring as hell but if you can't stand that, then you don't become a corpus linguist.

Maybe as a side remark only, I mean, slightly polemic, but I do think it's necessary sometimes. I mean, of course, this kind of stuff [pointing to the lower part on Figure 15] of course] takes weeks sometimes, I mean, reading all those examples and annotating them. And a lot of times, I get these weird looks when I tell people, there were sixteen thousand or nine thousand items or something like that. And then you go to conferences and you see these talks where people present two hundred examples. That's not really enough. In other sciences, [such as] in archaeology, people dig up some stuff for three months in the desert. So why, as a corpus linguist, would you think you're done after an afternoon of coding? It just doesn't make sense. So I do think getting your hands dirty with a lot of actual data is important. And that is not something that you can just breeze through, because it's so much more fun to run the nice graphs or something.

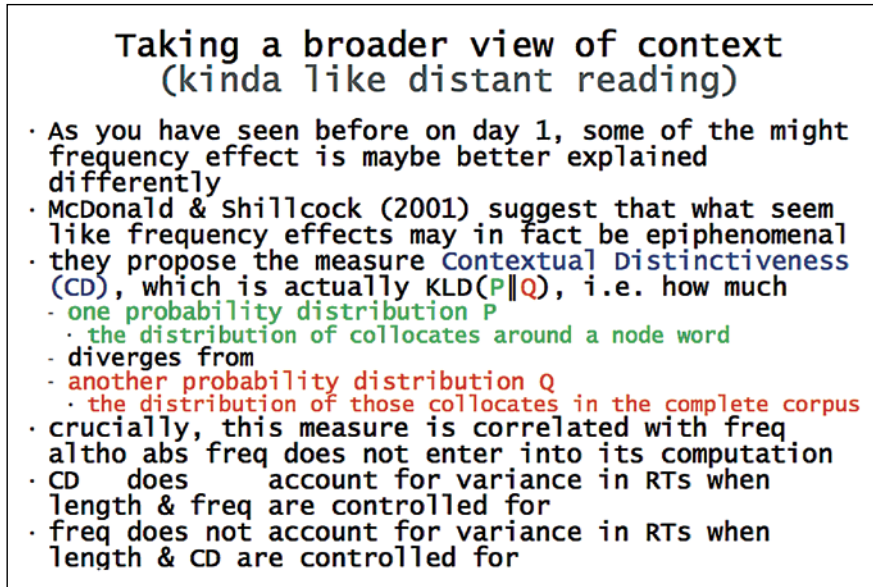


FIGURE 6

So the broader view of context here is quite different. The narrower view of context is the one kind of what you've seen a little bit now. You read every single concordance line, you annotate it for features of noun phrases, or verb phrases, whatever. But there is also a potential weakness to that and that is the fact that you actually are kind of drowning in details. I mean, every little word might have something of relevance to offer for the use here. Every construction, every syntactic pattern might contribute something, if you have a phonologically-annotated corpus, it's even richer than that. So it's very easy to read hundreds of lines, every line one by one, focus on exactly what's going on there. And I'm not saying it's a bad thing—all I'm saying is that it's incomplete because there is this thing that in digital humanities now might be called or has been called *distant reading*, so basically taking a step back and see what broader patterns are there? So patterns that transcend every single example, that basically actually even transcend maybe hundreds of examples or something like that.

As I mentioned earlier, some of these effects that translate into frequencies or something can perhaps be better explained in a different way. And so one really cool study is one that I mentioned briefly before and now I'm going to discuss it in more detail. This is a really cool paper by McDonald & Shillcock (2001) who suggested, as now a few others have, that frequency effects are epiphenomenal and they proposed this Contextual Distinctiveness (CD) measure

Taking a broader view of context (kinda like distant reading)						
words	a	b	c	d	e	Σ
P: Freqs of words around <i>state</i>	8	1	2	1	3	15
Q: Freqs of words in corpus	10	15	5	10	15	55
P: % of words around <i>state</i>	0.5333	0.0667	0.1333	0.0667	0.2	1
Q: % of words in corpus	0.1818	0.2727	0.0909	0.1818	0.2727	1

FIGURE 7

that I alluded to before. And the way this is operationalized is actually with a very famous information-theoretic measure, namely a measure that measures the distributional differences between two probability distributions. So what that means is that it measures how much one probability distribution, which could be something like this, the distribution of collocates around a word of interest. So you do a study, and you look for a word, let's say the word is *state*. Then you have the concordances. You have that concordance, let's say you have three words before and three words after. Let's say you have 100 examples of *state*. So that means, you have 300 words in front of it and 300 words after it. And then you generate a frequency table of that, which might look like this [referring to the table in Figure 7].

So the word *a*, occurs eight times around *state*, the word *b* just one time, the word *c* occurred two times around *state* ... so you have the frequency distribution of all the collocates. Then, this measure measures how much that distribution is different from the distribution of the same collocates but now not around *state* but everywhere. So what that expresses then is basically, how much does the presence of this word [*state*] affect the frequencies of things around it? So it's half-way decontextualized, because it doesn't say, for instance, whether this word occurs before *state* or after or directly next to it, or a little bit away from it. It doesn't use much contextual information, but it uses some, namely how much does the word *state* when you put it somewhere, sort of affects everything around it.

And I'm not going to show you exactly how this is computed here, because that's beside the point. But I just want to show you the general approach here. So these are the frequencies of words around *state*, and I was restricting it here for the fun of it to fifteen examples, fifteen collocates. And so what you compute then is basically out of all the words around *state*, how much is this in percent? Fifty-three. Eight out of fifteen is fifty-three, one out of fifteen is that (the



frequency of word *b*); two out of fifteen is that (the frequency of word *c*) and so on. And then you look at the frequencies of the same five words here in the corpus as a whole. And so then you might find that there are some words that behave very similarly and some words that behave very different. So when *state* is there, this one is not, just a single time, but that word is actually relatively frequent in general. So that means *state* suppresses that, so to speak. Same here: The word *b* shows up around *state* only a single time, but it's relatively frequent in the corpus so there's a huge discrepancy in terms of these percentages. And so that's what that measure—in a way that again, I'm not discussing here—but that's what that measure expresses. And like I said before, this measure is correlated with frequency but the absolute frequencies don't actually enter into the computation. Because the computation is done on the basis of these percentages: It doesn't matter whether this is eight and ten or eighty and one hundred.

And again, it accounts for variance in reaction times when length and frequency are controlled for, whereas frequency does not [account for variance in reaction times] when this (length and contextual distinctiveness) is controlled for. So this is basically a view of context that sort of is taking the step back from every individual line and looking at what kind of patterns are there across everything. And in a way, I hope it's obvious but just to make it very clear: This is also something that you cannot see without quantitative methods. You can't have an Excel spreadsheet with a thousand items of something. And then you can't eyeball the context in search for some patterns. That's not

**Taking a broader view of context**  
(kinda like distant reading)

As mentioned earlier, this would have implications for models of psycholinguistic models of lexical storage & access

- if freq-as-rep has a (bigger) impact on lexical access, this would be compatible with a model in which
  - activation of nodes can over time lead them increasing their resting level of activation
  - which means they are then more likely to fire again (more quickly)
  - (even this would still require dispersion, given its indisputable effect on learning)
- if freq-as-rep has little to no impact on lexical access, this would require a different approach
  - nodes become more likely to activate not because they are firing a lot
  - nodes become more likely to activate because they are activated from many different other nodes
- oversimplifying, the difference would be one of 'entrenchment vs connectivity'

FIGURE 8



going to happen. You do need to sit down and do some sort of math in order to figure that out.

And again, as mentioned before, at one point, this does have implications for psycholinguistic models of lexical access, lexical storage, and things like that because like I said before, if frequency-as-a-repetition-effect was more important, then we want to assume a psycholinguistic model where nodes are activated time and time again and their resting level of activation makes them potentially more likely to be activated, although we would still need dispersion for that. But if frequency-as-repetition has no impact, and in fact, it's contextual things like the ones I just discussed, then we need a psycholinguistic production or storage model that is based a lot on connections from words to others. It's not so much the resting level of a single node—it's how much does a word connect to everything else and basically gets, for instance, partial activation from those nodes. So in a sense and of course, this is oversimplifying a lot. I'm aware of that, but the difference would be one to say, are we interested in entrenchment of a single node as a function of prior activation or are we assuming what is actually the cause of something like connectivity, so the degree to which a word in a semantic network connects to many other words at the same time [[might be the cause]].

Again, I don't know what the answer is. All I'm saying is much of usage-based linguistics has been running with this for twenty, thirty years now, when in fact by now, there's empirical evidence accumulating that maybe this is actually playing a bigger role but very few people are talking about this.

### Taking a narrower view of context (similar to the contingency one)

On day 1, we mentioned type-token ratios as a means to summarize a distribution (and mentioned alternative measures from SLA/LCR)

- this can apply to words etc in a corpus as a whole
  - this can apply to, say, words in a constructional slot
- then we said that entropies ( $H/H_{\text{norm}}$ ) are a different way to quantify the evenness of a distribution
- you've already seen this plot,  $H$  has something to offer beyond type freq, token freq, & their ratio
  - $H$  can be computed over the frequencies with which
    - verbs a-h occur in constructions m-x
    - constructions i-p take verbs t-z
  - note:  $H$  can be related - if you want to - to dispersion: you quantify how the verbs are dispersed over the constructions
  - $H$  is related to Zipfianness of distributions: the more extremely Zipfian, the lower  $H$

FIGURE 9

So, that's one way to look at context. Another one is, again, a little bit narrower. And it goes back to something we talked about at the very first talk and the second one as well, namely, this idea that we can use, type-token ratios to quantify a certain distribution. I told you at the beginning that these are very dependent on sample size. I suggested some alternatives. And the crucial thing is that, of course, we can apply something like type-token ratios or lexical diversity both to a corpus as a whole, or to whole files in a corpus, but from a linguistic perspective, maybe more interestingly to words in a constructional slot. And then what we said before is that something like entropy is a way to quantify a distribution that actually supersedes or goes beyond what type-token ratios have to offer.

I already showed this plot [referring to the bar graph in Figure 10] before. So this is only really for those people who haven't been here during the first two talks, the first talks that mentioned that. So again, these are the verb forms of the verb *give*. We have a certain number of types, namely *give*, *gives*, *giving*, *gave*, and *given*, so five different types. We have 1,229 tokens altogether. And these are all possible distributions that you can get with that number of types, with that number of tokens and that type-token ratio. And the plug that I made for entropy earlier was that these entropy calculations here can distinguish between these kinds of distributions.

Now you can compute these over, like I said, verbs that occur in a construction. But interestingly, you can also compute these, for instance, on constructions taking certain verbs. So what that means is what we can express with entropy is, or what we can characterize with it is, sort of the constructional diversity of different verbs. How polygamous, if you will, are certain verbs in allowing for co-occurring with different constructions? And at the same time here, we can quantify how permissive or promiscuous constructions are in allowing different verbs to occupy their verb slots. And this kind of measure, entropy, is actually correlated with entropy in a certain way. Let me show you this here.

So here we have a bunch of type-token distributions [referring to the eight graphs on the right part in Figure 10]. The numbers of types are here on the x-axis in each of these panels and the number of token frequencies are up here on each y-axis. And the main point to show here only is that as the distribution becomes more and more Zipfian, more and more skewed towards a small number of very frequent words and a large number of extremely infrequent words the entropy goes down. So again, forget about the equations or anything. The main point is to show that the shape of this curve is reflected in that number, and that is relevant. Why? Because by now we have a whole bunch of studies that support that the role of entropy is quite important.

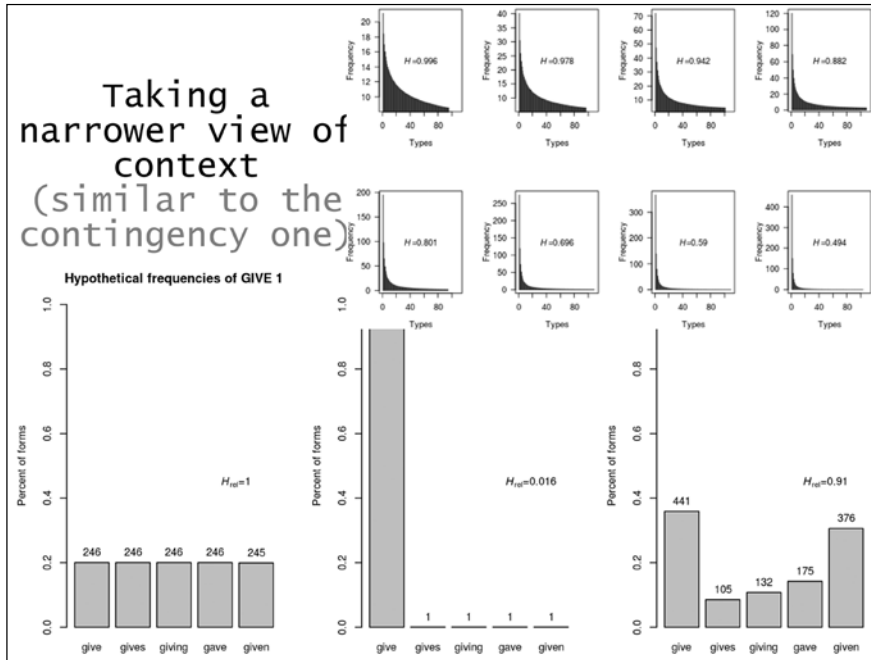


FIGURE 10

And we care because ...?  
 $H/H_{rel}$  and experimental data

Casenhiser & Goldberg (2005) find that children and adults learn a new construction better from skewed than from balanced exposure (5 verb types, 16 tokens)

- skewed condition: 8-2-2-2-2 ( $H_{rel}=0.86$ )
- balanced condition: 4-4-4-2-2 ( $H_{rel}=0.97$ )

$$H_{rel} = -\sum_{i=1}^n p_i \log p_i \cdot \frac{1}{\log n}$$

Boyd & Goldberg (2011: exp. 2-3) show that speakers

- learn to not use 4 novel a-adjectives prenominal from only 3 exposures to 2 of these adjectives in a preempting relative clause context
- distinguish preemptive from pseudo-preemptive contexts

we also know from non-linguistic categorization that categories with lower member type frequencies, lower entropies, and much exposure to the prototype are learned better than more representative categories

"in category learning in general a centred, or low variance, category is easier to learn" (Bybee 2010:89), and words in a cx slot are a category

FIGURE 11

So here's an earlier study by Casenhiser & Goldberg (2005) from a context of language acquisition. They exposed both children and adults to nonce constructions, so constructions they didn't know yet, because actually they don't exist in that language. And what they wanted to test is basically to what degree the shape of the input distribution facilitates, or makes harder, the learning of those constructions. So they gave five verb types distributed over sixteen tokens to children and adults to learn them and they did that in two conditions: One is a skewed condition and one is the balanced condition. What does that mean? It means this: In the balanced condition, the five verb types were relatively equally frequent over the sixteen tokens: It was 4–4–4–2–2. In the skewed condition, there was one strongly associated verb, sort of the path-breaking verb, and everything was much rarer. And then they showed that the skewed condition facilitates learning compared to the balanced condition. What they didn't do at the time is they didn't even think about entropy because back then, it wasn't a big deal. But if you do compute the entropy, you see that this one here has a lower entropy value, and entropy is often considered to be a measure of uncertainty. And so if the uncertainty here is lower, then the information communicated by this occurrence here is much higher, so that would be one way to explain why in their experiment, in spite of the same type and token frequency and type-token ratio, it is entropy that perfectly predicts that this one [skewed condition] should be learned better than that one [balanced condition].

In another study by Adele, this time with Jeremy Boyd (Boyd & Goldberg 2011: exp. 2–3), they showed that speakers learn to not use *a-adjectives* [namely,] adjectives that begin with *a*, pre-nominally from only three exposures to two of these adjectives in a preempting relative clause context. And the cool thing there was that the speakers in the experiment, they even distinguished preemptive contexts from pseudo-preemptive contexts. So under the hood, what the learners did is they realized “here I shouldn't be using this kind of adjective for this reason and here I shouldn't be using it for some other reason”. So they were balancing different kinds of conditions in the distributions that they were exposed to. And of course, we know from all other kinds of areas of categorization, such as non-linguistic categorization, that categories with lower member type frequencies, lower entropies, and more exposure to the prototype are learned better. So there's really a whole lot of experimental literature. I mean, these are just some examples that already shows that this is the case. And even someone who's actually not that experimental, someone like John Bybee (2010:89), even she says, “in category learning in general, a centered or low variance category is easier to learn”. And words in a constructional slot are a category. So this is the first set of examples, you know, for why we would care. There's experimental evidence that shows entropy plays a role.

## And we care because ...? $H/H_{rel}$ and observational data

- Linzen & Jaeger (2015) find that the  $H$  reduction of potential parse completions is correlated w/ **reading times of sentences** involving the DO/SC alternation
  - *Worf accepted Picard was right* ( $\rightarrow$  lower  $H$  of complem.)
  - *Worf forgot Picard was right* ( $\rightarrow$  higher  $H$  of complem.)
- Blumenthal-Dramé (2016:500):  $H$  of verbs' subcategorization frames correlates w/ **activity in the anterior temporal lobe** 200–300 ms after the stimulus
- Lester & Moscoso del Prado (2017) find that  $H$ s of syntactic distributions affect
  - **response times of Ns in isolation**
  - **the ordering in coordinate Nps**
  - "words are finely articulated syntactic entities whose history of use partially determines how efficiently they are produced [...] Perhaps words and syntactic structures are much more tightly linked than is typically acknowledged"

FIGURE 12

But we also have observational evidence from corpus data. So for instance, in a study from 2015, Tal Linzen & Florian Jaeger find that the entropy reduction of potential parse completions is related with reading times in an alternation that is called the direct-object subordinate-clause alternation. Here's an example that hopefully shows what this does. So here's one sentence: *Worf accepted Picard was right*. And here's another. *Worf forgot Picard was right*. So what this means here is that the word *forget* has a wider range of syntactic patterns that it co-occurs with: You can use it intransitively, you can say, *I just forgot*; you can use it transitively, *I forgot my keys*; you can use it with the *to*-clauses, *I forgot to bring my keys*; you can use it with a sentential complement here like, *I forgot Picard was right*.... So it takes a whole lot of different things—*accept* doesn't. It takes way fewer different subcategorization patterns. So that means when a subject reads the sentence, *Worf accepted*, by the time that the reader recognized *accepted*, the uncertainty of the syntactic structures that might follow has been reduced quite a bit. Because there's only a handful of things that can come after this [*accepted*] whereas this [*forgot*] leaves a whole universe of choices to open. And so that the lower entropy here makes people read this faster.

Then, in a recent article in the special issue of *Cognitive Linguistics*, Blumenthal-Dramé (2016:500) showed that the entropy of verbs' subcategorization frames, so something like this here in fact, correlates with measurable activity in the brain in a certain time window after the stimulus has been

presented. And then a former Ph.D. student of mine and a former colleague of mine, they (Lester & Moscoso del Prado, 2017) did a whole bunch of studies looking into entropy, finding really cool stuff. One of them is this: entropies of syntactic distributions of nouns, they affect the response times towards nouns, of nouns, in isolation and the ordering of those nouns in coordinate NPs. I'm not even sure, I probably had to read this multiple times to figure out how great this is. So entropies of syntactic distributions, so how many different syntactic contexts does a noun occur in?, that has an impact on how fast you read that noun even when you *don't* get it in a context. That's the cool thing about that. Because it suggests, this is a quotation from them, "words are finely articulated syntactic entities [and now] whose history of use partially determines how efficiently they are produced". So the idea being that we carry around with each of us a, somewhat spotty, I guess, record of how and what kind of syntactic constructions nouns were used and that even affects how we process these nouns in isolation.

Some other things that they also found are that "words occurring in similar distributions of syntactic constructions prime each other independently of orthographic and semantic similarity in a visual lexical decision task". Again, I mean, you need to let that sink for a moment: So what they found is that if nouns have similar ways of being distributed, then they prime each other even if you control for whether the word looks like that orthographic[[ally]] or has the same similar meaning. So there's a trace of the distributional behavior in the past: A verb or a noun that is used with this construction this much, and this one that much, and then one like this, is going to prime another noun that has a similar spike in its distribution. So, pretty cool.

And then finally, from their work, they show that the phonetic duration of nouns, so how long does it take to pronounce that noun, is correlated with the diversity of a noun's distribution across its syntactic relations. So the more diversely nouns are used, the more that affects how long you take to pronounce them. So a direct effect of syntactic history on articulation. A whole bunch of things, where we not only see that entropy plays a role, but also how much, in a sense, usage-based linguists should be celebrating this all day long, because it makes such a strong statement about how we keep track of the history of use of things and it affects stuff that are going on right now and will be going on in the future. So really extremely cool stuff that they did there.

Now, how can we compute something like this? This is where the bad news part begins. It's not straightforward, not because the math isn't straightforward—for once, that is actually really straightforward—but we don't have those data a lot of times. What we need is essentially the beginning of something like this: A table that contains all sorts of verb lemmas and whose

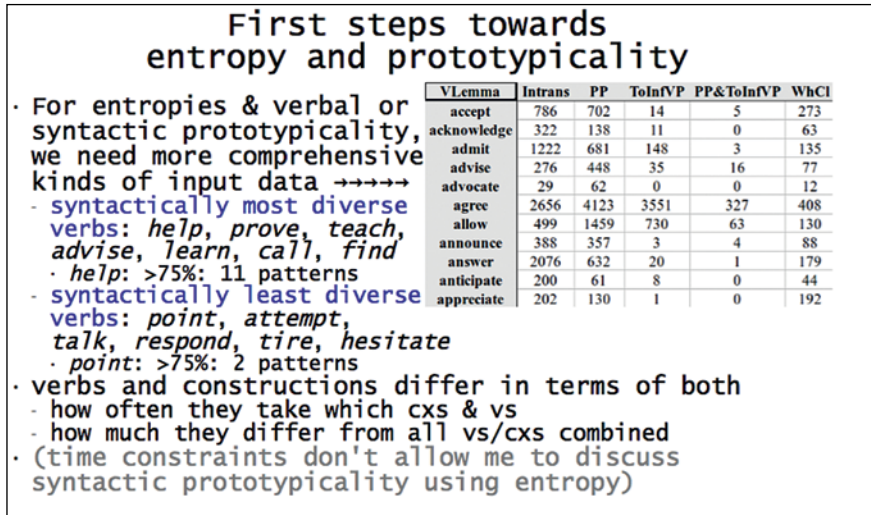


FIGURE 13

rows are all sorts of constructions at whatever level of annotation you think is feasible, and then you have co-occurrence frequencies. Here, for instance, in this data set, *agree* is used a lot intransitively like when you just say, *Yeah, I agree*. It's used a lot with prepositional phrases and I bet a ton of those are *I agree with you, I agree with him* or *I agree* plus *to* infinitive verb phrase *I agree to make this talk available* or something like that. And other verbs are obviously used more rarely, but also very differently. Let's look at this: I mean, *allow*, the ratio of this [499] to this [1459] is one to three but for *announce*, it's more like one [388] to one [357]. And for *answer*, it's more like three [2076] to one [632]. So every one of these verbs has a different line connecting the percentages of its use in different kinds of constructions.

But of course, the pain in the ass is that we don't have this. And we can get this from parsed corpora, but they will always come with a certain degree of imprecision or errors in the data. So the more we try to do this, the harder it'll get. But with stuff like that, for instance, we can look at things like this: So which verbs are syntactically most diverse? And if you run this on this dataset, then you come up with this pattern here. This actually contains data on more than two hundred verbs if I remember correctly: *Help* is the syntactically most diverse verb in this dataset. If you want to cover approximately 75% of the uses of *help*, you need to look at eleven different subcategorization patterns because it goes with so many and not particularly frequently, but it's very widely used. The opposite would be the syntactically least diverse verb, that would be *point* and then *attempt*. If you want to cover 75% of the use of *point*, you only need to look at two patterns and then you already have that verb nearly covered.

And so that kind of stuff would of course be very interesting, for instance, from a perspective of acquisition studies, one would assume, for instance, that, obviously, [for] verbs that are more diverse, if the child wants to arrive at a full command of what that verb would do, obviously this would take much more here because it's so much more diversely used.

So both verbs and constructions differ in terms of how often they can take certain constructions and certain verbs respectively and how much they differ from everything else. For instance, it's possible for every verb to compute how much it differs from all other words. And again, how would that not be interesting for acquisition, if you can say, "okay, this is a verb that has a very atypical behavior, given everything else we've seen the child use in the past". So, that might inform predictions about if the child hears this a number of times, will it pick it up or not depending on, for instance, how large the difference between any one verb is and everything else the child already knows. I think there's very clear connections to be made.

Now when it comes to this kind of stuff, we can by now even go so far as to find cases where the expectations that the speaker has, given things they have heard or have produced, play a role, in particularly for learning. We've already seen this in the example of the entropy reduction, like *accept* makes you faster because the number of potential continuation is smaller. But it goes even further than that. So one way to look at this would be one that has been looked at a lot by some people from Rochester but also Nick Ellis again, so the idea is that learning is driven by prediction errors. What does that mean? It means this: Namely, we learn more from the surprise that comes with you expecting *a* and you're getting *b*, than when you're expecting *a* and you're getting *a*. So unexpected things make your cognitive system perk up for a moment and be

the amount of learning induced from  
an experience of a construction depends  
upon the salience of the form  
(i.e., how much it stands out  
relative to its context) and the  
importance of understanding it correctly  
(Rescorla & Wagner 1972, Ellis 2006)

Ellis, Römer, & O'Donnell (2016:47)

FIGURE 14



## In fact, even our expectations play a big role when it comes to learning

- Some contemporary learning theories hold that learning is driven by prediction errors
- we learn more from the surprise that comes when our predictions are incorrect than when our predictions are confirmed (Rescorla & Wagner 1972, Rumelhart, Hinton, & Williams 1986, Wills 2009, Clark 2013)
- there is increasing evidence for surprisal-driven language processing & acquisition (Demberg & Keller 2008, Jaeger & Snider 2013, Pickering & Garrod 2013, Smith & Levy 2013)
- surprisal is often operationalized as  $-\log_2 p$ , ie
  - the less likely something is, the more we are surprised
  - the more likely something is, the less we are surprised
  - $p$  is typically a conditional probability, e.g.  $p(v|cx)$
- the surprisal of a word in a sentential context is the probability mass of the analyses not consistent w/ it (Ellis, Römer, & O'Donnell 2016, Demberg & Keller 2008)

FIGURE 15

like, “oh, well, okay, I didn’t see that one coming” and that makes you learn better.

And by now, this unexpectedness, which is called surprisal, and I’ve shown you the formula one time before, has been shown to have quite some impact on both processing and acquisition. And the cool thing about this one at least is that it’s extremely easy to measure: You just take the probability of whatever you’re interested in and you take the log to the base of two and make it negative. What that means is the more unexpected something is, which means the smaller the probability of something, the less you expect it, and the higher this value will be, because of the logging and the negative. If something, on the other hand, is very likely, then this will turn that surprisal value into something very low. If something is very likely, you’re expecting it, if it comes, you’re not surprised. So there’s an inverse relationship between probability and surprisal. Basically what I just said here and typically this is not just an overall probability, but a conditional probability: How likely is this verb given that construction? How likely is this sentence continuation after you heard *accept* as opposed to *forgot* or something like that?

One very nice way to define this is this: “The surprisal of a word in a sentential context is the probability mass of the analyses not consistent with it”. So after every little bit of input that you’re getting from the speaker, you’re

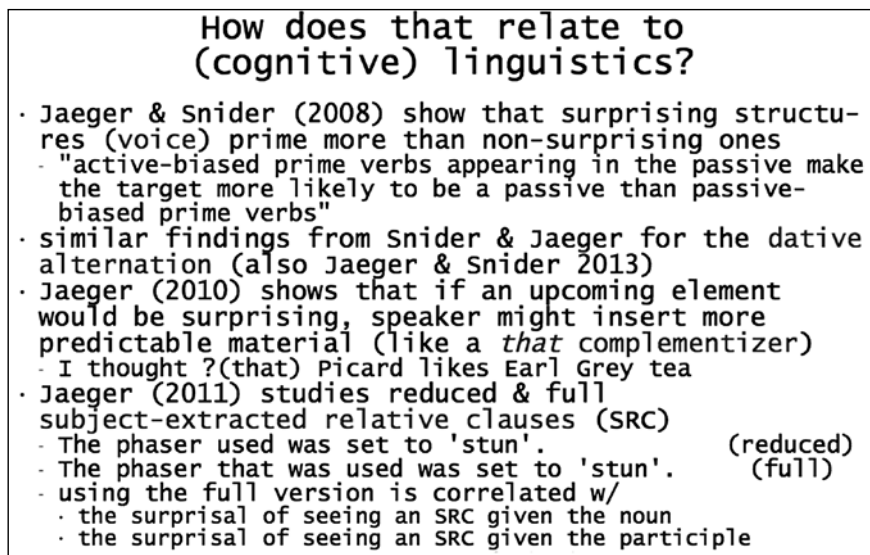


FIGURE 16

processing it and your brain tries to anticipate what's going to come next. And if something is very surprising, then most of the things you were considering were not compatible with what actually happened. That's kind of the idea here.

So how does this work? One manifestation of it is actually in priming. Remember: priming was this tendency to reuse a construction or a syntactic structure you've used before. And I think the example that I gave you was that if you see a transitive scenario and you describe it with a passive sentence, then you're more likely to describe the next scenario also with a passive sentence as opposed to the default active.

And so what they found is, so this is a little convoluted and hard to process: active-biased prime verbs—that means is verbs that like to be in the active voice—appearing in the passive, make the target more likely to be a passive than passive-biased prime verbs. So if you have a verb that likes active very strongly, but you use it in the passive, then that's surprising. That gets noticed and makes it even more likely that you're using it in passive next time around because it didn't match your expectation, and they had similar findings for the dative alternation and also for *that*-complementation.

So what people do is, if an upcoming element in what they're going to say is surprising, then they smooth it over by inserting something that is predictable. For instance, that's one reason why sometimes people insert a *that* complementizer and sometimes they don't. *I think he's a good colleague* would be more

## And now for a word of caution ...

- We have seen that
  - corpus data have a lot to offer that is routinely not employed in cognitive/usage-based linguistics
  - these data require us as linguists to correlate corpus frequencies and other statistics w/ other cognitive, linguistic, contextual, or any other annotated data
- but usage-based linguists makes the claim that this is actually what speakers do the whole time, too:
- "we need to conceive of grammar as based on constructions & as having an exemplar representation in which specific instances of use affect representation. (Bybee 2006:714)
- I am highly sympathetic towards that claim ...
- but it's also a rather strong claim/assumption, esp.
  - in the exact form that it is made
  - in combination w/ other claims/assumptions usage-based linguistics are making
- how generally are we monitoring correlations?

FIGURE 17

likely than *I think that he's a good colleague*. Because after *I think*, the word *he* is relatively expectable. That's going to happen a lot of times. But if you say, *I think the terrible dinosaur, the terrible dinosaur* is not particularly expectable, so people might insert the *that* there, which comes with very little processing load, but explicitly marks syntactic structure so it makes the processing of the upcoming material easier. So basically inserting *that*, there is a way for the speaker to minimize surprise for the hearer. Okay, that's the idea.

We've seen now a variety of ways in which context can be useful. I mean you can have the fine-grained reading of concordances; you can have the distant-reading approach with this contextual distinctiveness; we might look at certain slots and compute entropy and surprisal and all sorts of things like that.

But what I want to turn to now is basically that, to the more general question, taking a step back if you will, is how much of this context we actually notice. Just because there's something in the context that as a corpus linguist we can find, if we look at a concordance twenty hours a day, that doesn't mean that speakers notice. Corpus data offer a lot, but at the same time they require us, and this brings us back to the first section and the second one, as linguists, we correlate corpus frequencies and other kinds of statistics with cognitive, linguistic, contextual, and other kinds of annotated data. I think that much is uncontroversial.

But the thing is that usage-based linguistics makes the claim implicitly—actually no, explicitly, but it doesn't commit to how exactly this happens a lot of times—[that] this is what speakers do the whole time. The idea is [that for]

every instance we hear, we place it somewhere in multi-dimensional space and then like Arie already said, categories are maybe not formed, but they emerge from exemplar clouds and stuff like that. Here's a quote from Bybee (2006:714): "We need to conceive of grammar as based on constructions [ok, sure] and as having an exemplar representation in which specific instances of use affect representation". "Specific instances of use", which of course I uttered in a certain context with certain functions and so on "affect representation". And I like that approach. I, for a long time, have been using sort of exemplar-based kind of language in analyses that I've proposed myself as well. But I would want you to realize that this is a very strong claim to make or a very strong assumption, especially in the exact form that it is made a lot of times and in combination with a lot of other claims that usage-based linguists are often making. So the question I'm trying to raise here sort of is: how generally really are we monitoring co-occurrence relationships or correlations in general? So let's revisit some of these claims that as cognitive linguists, we always uphold as our main working assumptions.

One would be this, "the structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes" (Clay et al. 2009). How can we not like that? That sounds exactly like what we would want. What are the general mechanisms that are proposed here? One, for instance, is especially Joan Bybee (2010:79) being a strong advocate of this, that "many of the things that happen in language are domain-general". So in explicit contradistinction to the generative approach with its modularity view of things, here the idea is, let's push this idea of domain-general cognitive mechanisms

### The Usage-Based Theory of Language and its assumptions

- Main tenet: "the structures of language emerge from interrelated patterns of experience, social interaction, and cognitive processes" (CB et al. '09)
- relevant principles/mechanisms
  - domain-general cognition
    - incl. chunking, categorization, cross-modal association, analogy, and rich memory storage (JB 2010:79)
  - no a priori distinction between different levels of (linguistic) structure: 'everything's a construction'
  - mental representations of language change throughout life
  - rich memory of specific experiences
    - incl. linguistic, non-linguistic, contextual, & inferential information, from which structure is extracted
  - note: this is not an incidental claim that could be withdrawn while leaving the rest of the UBTL intact – the UBTL leaves it open how much/what kind of non-linguistic information is stored and how

FIGURE 18

as far as we can. And those include things like chunking (maybe giving rise to constituent structure), categorization, cross-modal associations, so connections we form between different sensory modes, analogy, and obviously, if you assume an exemplar-based approach, you have to make the assumption that there's very rich memory storage. Because how otherwise do you have enough exemplars in your memory to form categories over?

Also, again pretty uncontroversial I think in usage-based circles, there's no a priori distinction between levels of linguistic structure. "Everything is a construction". Morphemes, words, multi-word expressions, whole proverbs can be stored as a whole unit so everything's at some level of generality in the construction. Then, mental representations of language change throughout life. Even when you are like two minutes before death, you can still learn a new word. Maybe that's not what you want to do at that point in time, but it would be possible. So obviously we're processing these things all the time in every single usage event has an effect.

This one is important, "rich memory of specific experiences", which if you look at the relevant literature, it includes linguistic, non-linguistic, contextual information; it includes inferential information—again something that was mentioned a lot of times in your talks—and from all of those pieces of information, structure is extracted. Now I want to be a little critical here, this is not an incidental claim. Especially rich memory here. You can't say, "okay, I'll buy everything else", but this one, "okay, I give that one up." No, I mean, if you assume a usage-based approach on the basis of exemplar models, then you do need to commit to something like this, because otherwise there's no mechanism that provides all those rich memory representations from which you extract categories. This is a central cornerstone, even if it's not always explicitly flagged as/like that.

### The Usage-Based Theory of Language and its assumptions

- There seems to be evidence for each component of this theory
- but note
  - the posited processes are extremely general in nature
  - that also means the UBTL explanation for learning linguistic structure amounts to a description of how we learn *any* structure (any statistically reliable pattern in the world)
  - the extant work leaves considerable uncertainty about just how broad/ly they may be/apply

FIGURE 19

If you look at all these components, it seems like there's evidence for every one of them. Again, I'm actually very much in favor of exactly this view. I'm not trying to tear down the whole building here. But the posited processes are extremely general in nature precisely because they're supposed to be domain-general, cut across language and other kinds of cognitive modules—to use the *m*-word—which of course means that the usage-based theory of language explanation for learning linguistic structure pretty much amounts to a description about how we learn *any* kind of structure. Because exemplar-based learning is something that can be applied to pretty much anything you learn, anything episodic, anything procedural, and anything linguistic. The posited mechanism works for all of those. But the relevant work in this context actually leaves considerable uncertainty about how far you can push that.

Now for instance, if we look at domain-general cognition, something like chunking. I'm assuming probably everyone knows here what that is: It's the process by which sequences of units that are used together cohere to form a more complex unit. Chunking is one way in which we abstract, for instance, multi-word expressions. What did I use before as an example in the week? Yeah, there's not a lot happening after *in spite* that is not *of*. So after the thirteen thousandth time, we just say, maybe *in spite of* is just the unit and we make that a chunk because it works. And of course, learning statistical regularities is not modality-specific. The fact that things could occur together, you don't just do that in language, you do that in everything. It's the reason why you turn left before you cross the street. You look because a lot of times someone there was coming and it's a good idea to then check.

**The UBTL involves  
domain-general cognition**

- **Domain-general cognition**
  - **chunking**: the process by which sequences of units that are used together cohere to form more complex units
    - learning statistical regularities is not modality-specific
  - **cross-modal association**: forming associative links between disparate aspects of experience (Law of Contig)
    - weird word order studies, simulation semantics, modeling, ...
    - this is not unlike Pavlovian conditioning (Rescorla 1988)
      - how do these associations really work?
      - how much extralinguistic context is available for association? "exemplars are tagged for their contextual associations, *both linguistic and extra-linguistic*" (Bybee & Hopper 2001)
  - chunking involves units of the same kind & cross-modal association involves units of different kinds so they may instantiate the same process

FIGURE 20



## The UBTL involves domain-general cognition

- **Domain-general cognition**
  - **rich memory storage**
    - exemplar representations are "rich memory representations; they contain, at least potentially, all the information a language user can perceive in a linguistic experience. This information consists of phonetic detail, including redundant and variable features, the lexical items and constructions used, the meaning, inferences made from this meaning, and from the context, and properties of the social, physical and linguistic context." (Bybee 2010:14)
  - **categorization**: linguistic knowledge is built up by recognizing patterns of similarity among exemplars
    - "[Language users] map similar tokens onto one another to establish exemplars and these exemplars group together to form categories that represent both the fixed and schematic slots in constructions. The meaning of a construction is also represented by a set of exemplars which are built up by accessing the meaning of the lexical items used plus the overall meaning in context." (Bybee 2010:26)

FIGURE 21

The same thing with cross-modal association. We know that we form links between something we hear and something we see or between a word, so something purely conceptual, and the phonological form with which it is created. And we know that from a ton of different studies, many of which actually are really not very much unlike Pavlovian conditioning (Rescorla 1988).

Now the question then is how do these associations work? And how much extra linguistic context is available for association? That's an important question here, because this is a direct quote here from Bybee & Hopper (2001): "exemplars are tagged for their contextual association [and now, what will be critical later is] both linguistic and extra-linguistic". So again, a very strong claim, very general claim, about how we do this.

Rich memory storage, the other big thing that I pointed out that is relevant for this. Again, a nice illustrative quote here from Bybee (2010:14):

Exemplar representations are rich memory representations. They contain at least potentially all the information a language user can perceive in a linguistic experience. [That's a lot!]. This information consists of phonetic detail including redundant and variable features, the lexical items and constructions used, the meaning, inferences made from this meaning and from the context [again:] and properties of the social, physical, and linguistic context." That is technical language for 'pretty much anything'.

## The UBTL involves a rejection of discrete levels of structure

- **Rejection of discrete levels of structure**
  - "all of the units of language – segments, phonemes, morphemes, words, phrases, constituents – can be arrived at by the simple categorization processes [in the exemplar model]" (Bybee & Beckner 2009)
  - "there is no analysis into units at any level or set of levels that will ever successfully and completely capture the realities of synchronic structure or provide a framework in which to capture language change" (Bybee & McClelland 2005)
  - once the above is accepted, one cannot counter adverse empirical findings by claiming that the rules are different for that structure – all domain-general mechanisms are fully available to operate on anything (including linguistic and non-linguistic aspects)

FIGURE 22

And so then, rejection of “discrete levels of structures”: “All of the units of language—segments, phonemes, morphemes [words, phrases, constructions]—can be arrived at by the simple categorization processes in the exemplar model”. So again, a claim about something extremely, extremely general. “There is no analysis into units at any level or set of levels that will ever successfully and completely capture the realities of synchronic structure”. Now the thing is once you accept all this, you basically shut off one nice excuse if things don’t pan out your way. Once you accept all that stuff, you cannot say, “this finding, and this finding, it didn’t work. Well, that’s because there the rules are different for that structure”. You cannot write a book with twenty pages on domain-general chunking and learning and categorization, and then say, “but for complement clauses, yea, no”. I mean, no allusion intended to anything you said.

## The UBTL assumes cognitive systems change throughout life

### Change throughout the lifespan

- "The usage-based model [...] proposes that even in adulthood our experiences with language continue to affect mental representations" (Bybee & Beckner 2009)
- "if usage is the basis of grammar and change in the grammar, then there is no a priori reason why change cannot occur over an adult's lifetime" (Bybee 2010:114)

FIGURE 23



## A particular kind of memory phenomenon: structural priming

- Structural priming/persistence (recall recency)
  - processing of a stimulus (prime) facilitates/makes more likely the processing of a related stimulus (target)
  - found in observational and experimental data
  - from comprehension and production to production
  - typical constructional examples
    - active vs. passive
    - ditransitive vs. prepositional dative
- structural priming is a natural fit for the UBTL: the implicit-learning account of structural priming essentially claims that it is a result of just the sort of pattern extraction posited by the UBTL (see, e.g., Chang et al. 2000, Savage 2006, Kaschak, Kutta, & Jones 2011, Rowland et al. 2012)

FIGURE 24

And finally, before we get to the experiment, “change throughout the lifespan”. The usage-based model commits you to that, “even in adulthood our experience with language continue to affect mental representations” (Bybee & Beckner 2009). “If usage is the basis of grammar and change in grammar, there is no a priori reason why change cannot occur over in adults’ lifetime” (Bybee 2010: 114). So basically meaning, even if you are an adult, you should still be picking up on correlations, domain-generally across everything you do.

So now let’s see. I talked to you about structural priming before. Again, the processing of a stimulus makes it more likely that you produce something related to that later as well. We find this in observational data and experimental data, we find it from comprehension and production to production, and a lot of times this is analyzed with these kinds of alternation-like examples, the voice alteration, and active versus passive—that’s the example I always gave you—but also something like ditransitive versus prepositional dative and so on. And in a way, the phenomenon of priming is a really good fit for the usage-based theory of language, because some people assume that priming is an implicit-learning phenomenon. It’s not there’s residual activation—it’s actually, you just learned that thing one more time, you just saw it one more time, and that has an impact on your system: That’s the natural usage-based theory of language explanation for why priming occurs, and there’s a ton of studies that have assumed exactly that.

Now there’s another interesting memory phenomenon that becomes relevant in this context and that is what is called context-dependent memory. That is a not completely but pretty robust finding that shows that if you learn things

## A particular kind of memory phenomenon: context-dependent memory

- Things learned in a particular context are more readily remembered when that context is reinstated
- the context can apparently be almost anything
  - music, one room vs. another, underwater vs. on land, ... (Smith 1985, Smith 1979, Godden & Baddeley 1975)
- most studies have focused on explicit memory but effects for implicit tasks have been found as well
  - homophone spelling (Smith et al. 1990, room context)
  - completing a maze (Parker et al. 2001, olfactory cont.)
  - word-fragment completion (Ball et al. 2010, olf. cont.)
  - picture naming (Horton 2007, person context)
- it seems to be the nearest thing to exemplar-like storage *outside of language*
  - it is sensitive to incidental contextual detail
  - it can apparently 'recognize' many kinds of contextual similarities
  - it can influence performance without awareness

FIGURE 25

in a particular context, then these things are more readily remembered if that context is reinstated. If you listen to classical music while you're studying vocabulary in a language that you're learning, then your memory for the vocabulary items you were learning will be better if, when you need them, you also play that same music. That's what context-dependent memory is about. Notice that in this example, for instance, the music has no bearing on the words. It's not—so it's classical music that's instrumental. It's not like the words that you're learning a part of the lyrics of a song or something. No, it's just that it's the same context.

In many of these studies on context-dependent memory, people have found contexts can be pretty much anything. It can be music, like the one example that I just gave you, it can actually be the room, one room versus another, it can be as freaky things as under water versus on land (Smith 1985, Smith 1979, Godden & Baddeley 1975). So if you learn vocabulary while under water, you will recall it better when you're under water. So maybe don't learn vocabulary under water, because I'm assuming you do not spend most of your time there.

Now, most of these studies have focused on explicit memory, but there are results for implicit tasks as well. So things like homophone spelling (Smith et al. 1990, room context), completing a maze (Parker et al. 2001, olfactory cont.), word-fragment completion (Ball et al. 2010, olf. cont.), picture naming (Horton 2007, person context), all sorts of things like that as well. So, in a sense, it seems to be the nearest thing to exemplar-like learning outside of language.

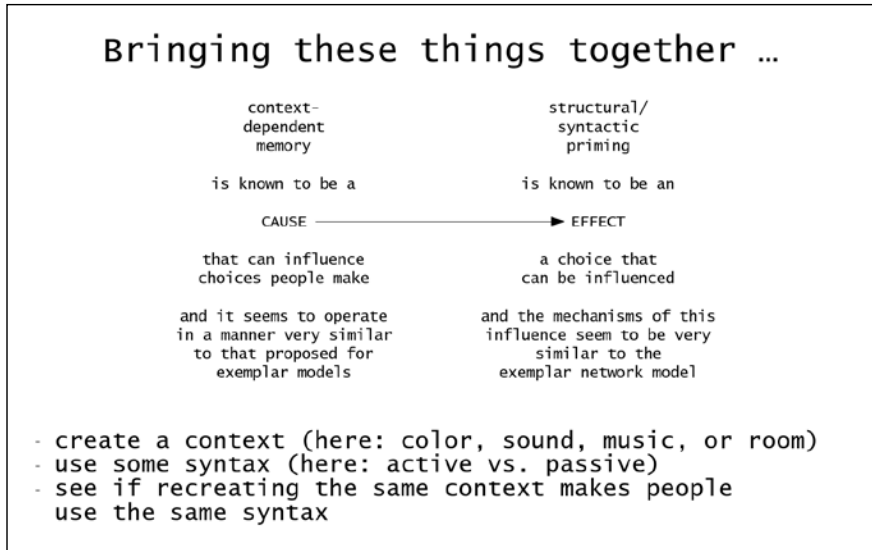


FIGURE 26

Because if you take the vocabulary example, in classical music, then classical music isn't part of your using language at the time but—remember all those Bybee's quotes and stuff—but you're picking up pretty much anything from the context. And in the usage-based theory of language, it informs your language processing, in context-dependent memory studies, it might affect your completing a maze, completing a word fragment, or recalling vocabulary. So this kind of stuff follows naturally from a usage-based perspective on language.

Let's bring these two things together, so what we end up with is we have context-dependent memory and structural/syntactic priming. In a sense, this is supposed to be the cause for the effect that we can observe in structural priming. The fact that we've just noticed the syntactic structure (even if we didn't actively monitor for it) makes us more likely to use something like this. This can influence choices people make and structural priming is all about the choices that people make. Do they choose an active or a passive? Do they choose a ditransitive or prepositional dative?

And then this seems to be operating in a way that is extremely close to what exemplar-based models and usage-based theory talk about. This then seems to be very much like what an exemplar network kind-of-approach talks about.

So Brendan [Barnwell 2014] did this nice experimental design. He created a context-dependent memory. He created a context, which in his set of

### Context-dependent memory and priming: the experiments

- Hypothesis: when presented a picture with a context associated with a particular voice, subjects will use that voice more often than when presented with the other context
- prime phase
  - subjects saw 24 sentences on a screen, one at a time, and read them aloud into a microphone
  - 12 actives and 12 passives were each coupled with a particular non-linguistic context (for each subject)

FIGURE 27

experiments was either color, sound, music or a physical location, a certain room. And then he had people use some syntax and his experiments that I'm talking about here used the active versus passive voice alternation. And then he checked "can you prime people to use actives or passives with this?". So if you recall vocabulary better under water when you studied it under water, then that means the context—submerged or not—gave you easier access to some words. Here the question is, if the color is repeated, if the music is repeated, if it's the same room—and room in particular has been shown to be have a huge impact—if it's the same room, can that prime constructions? Because this is the tricky thing here. Everything comes together. You might say, well, but vocabulary learning is different from choosing a syntactic construction. But in the usage-based theory, it's not. It's all constructions, right? So you can't say, oh "that only works for words and not for syntactic patterns". You can't say that if on the previous page you committed to a construction, it's constructions all the way around. So it should work.

The hypothesis was: when presented a picture with a context associated with a particular voice—active or passive—subjects will use that voice more often than when presented with the other context. We're looking at not structural priming, active primes actives, passive primes passives. We're looking at Mozart primes actives, jazz primes passives. Because usage-based theory says we're looking at all, we're taking it all in the whole context. So in the prime phase, subjects saw 24 sentences on a screen, one at a time, and then to make sure they process them correctly, they had to read them aloud into microphone. And so twelve actives and twelve passives, these 24, were each coupled with a particular non-linguistic context for each subject.

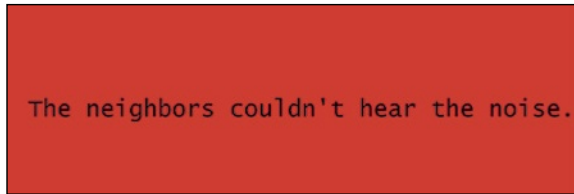


FIGURE 28

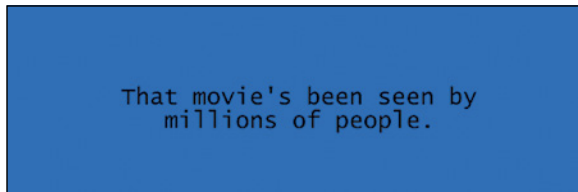


FIGURE 29

So this might be the active prime, so there's a red background and it is in active voice. And then the next one might be a passive sentence with a blue background.

That's the kind of stuff people saw, and similarly with the music, they might read a sentence—we have to play around with sequencing there, because obviously you can't give someone sentences to read for three seconds and play classical music and then three seconds later, the music switches to jazz, that would be totally distracting; we had to make some adjustments, but still it would work.

Then, in the test phase, he used, his dissertation, 24 pictures from one of the famous priming experiments in the 1980s. Each picture was accompanied by one of those originally non-linguistic contexts, so certain kind of music, location, color and the subjects were asked to describe the pictures into the microphone. So the pictures again show transitive scenarios and the question was, are they going to use active or passive or—annoyingly—something else. So those were the different options they could give.

And interestingly enough, it's hard to believe, but actually virtually no subject reported that he realized the patterning between the color and the voice. So no one came out of that experiment and in the debriefing phase said that "I saw these two colors, and one of them was always active". They just did not see that at all, which in a sense is cool because it means we are talking completely about implicit effects here. No one got it and therefore had to be discarded, because then it's concrete memory as opposed to implicit priming or learning.

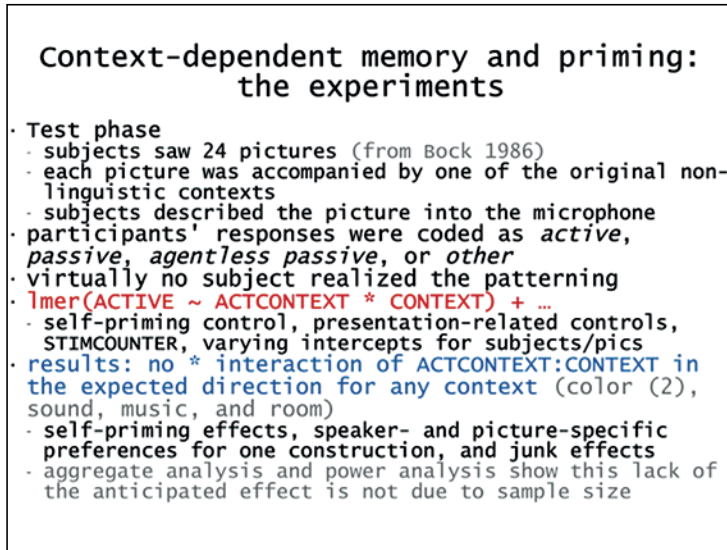


FIGURE 30

Then he did a statistical analysis of this. The dependent variable was whether people used active or not, and the main point of course was, was there an interaction between the contexts they learned to associate with actives and whether they got that one? So people should be using actives more in general than passives anyway. But if they saw a transitive scenario and they see it with a context that they have learnt to associate with active constructions, then that active frequency should be boosted.

So he used all the bells and whistles, I mean subject-specific controls and whatever, everything that you need to do. This is what he found: no significant interaction that showed that the context could prime the constructions. So the red and the blue thing, like in the example, did nothing. That was true for two color experiments and for sound and for music, and for room. In none of these things, context-dependent memory was found.

There were some other effects, so it's not like there was nothing. There were self-priming effects so if people use an active last time or a passive last time, they were more likely to do it again. There were effects that speakers have different preferences for the two constructions, certain pictures elicited more actives and others more passives. So I mean there was a whole lot of structure in the data, but not anything that is compatible with context-dependent memory of the type that was expected. And the fact that these effects were found also shows it's not like we had too few subjects in order to find anything significant.

## Interpretation

- The non-linguistic contexts were like a morpheme indicating voice – but subjects didn't recognize it
  - despite evidence from weird word order studies and artificial-language learning studies that suggest subjects *can* recognize such patterns
  - despite evidence from CDM studies that this kind of manipulation can induce people to do things
  - despite evidence from priming studies that people can be induced to use a particular voice
- the idea that people just notice patterns sounds nice – the problem is we have to
  - explain which patterns they pick up and which they don't
  - explore the constraints on the pattern-recognition process – type/token frequency (Bybee & Beckner 2009) is not sufficient
  - just because we can describe domain-general pattern recognition without reference to language doesn't mean it actually works across different domains

FIGURE 31

So slowly wrapping up. What does that mean? Well, the non-linguistic contexts were actually like a morpheme. Remember, in the training phase, they saw that every active sentence was in red with a red background. Every passive sentence was with a blue background. So they were, as regular as any actual active or passive morpheme in our language would be that would do that with a simple morpheme.

There was a completely exceptionless pattern, but the subjects didn't recognize it. They didn't recognize it explicitly for the debriefing, but apparently also not *implicitly* during the experimental trials, because then later there was no effect of that.

That's in spite of a whole bunch of studies and evidence that suggests they should have: Evidence from weird word order studies or artificial language learning. All these studies suggest that subjects *can* recognize such patterns. Despite evidence from all sorts of context-dependent memory studies that exactly the kind of manipulation that we did here can induce people to do things without their conscious recognition of the fact. And of course, despite evidence of the fact that we know very well, you can prime actives and passives. So it's also not like we choose a stupid construction that can't be primed. No, I mean there's decades of literature that show that you can. So the idea that people just notice patterns around them sounds really nice, but it's not specific



## Interim conclusions

- Now, some might say
  - "Duh! Of course, these experiments showed nothing. who gives a ... about background colors, ambient music, ...?"
  - the relevant criterion is that these 'ambient stimuli' do not reduce entropy – children/people have learned that background colors and ambient music etc. do not reduce uncertainty: backgrounds don't help you predict
    - words and their senses very much
    - constructions very much
    - meanings of sentences very much
    - inferences of sentences much
- all likely true ... but our current discussions of exemplar-based models etc. don't
  - discuss cue availability/recency, strength, validity enough ...
  - quantify the power of entrenchment against recency (would we get better results with new constructions?)
  - explain why CDM works w/ non-entropy reducing scenarios, but why we still can't create them with language

FIGURE 32

enough. nowhere near. We need to explain which patterns they pick up and, like here, which they don't.

And that means we need to formulate and explore constraints on this pattern recognition process, like type-token frequency or something like that as has been suggested, is not enough. And more generally, just because we can describe something like domain-general pattern recognition without reference to language doesn't mean that it works the same way across different domains. So it's a difference between us building up a model or a theory that can explain something in a certain way if we want to be cognitive realistic, than it has to happen that way.

Now you might say, "Duh! Of course, they showed nothing, these experiments, because speakers don't give a crap about the background colors and the ambient music while they talk". And in a sense, yea, the relevant criterion is something like, perhaps, entropy or salience. These "ambient stimuli", i.e. the background color of something printed on the screen or the location where you are physically sitting, they don't reduce entropy when it comes to understanding language. I mean, at no point in your life, probably have you learned that background colors make a difference to what you're hearing. If you're talking to someone in a brightly lit room and they say a certain sentence, then a week later, you see them again, and they say that very same sentence in a dark



room, you still know it means the same—most likely unless there's some like irony thing about the light going on or whatever. So backgrounds don't help you predict words on the senses or constructions or meanings of sentences or inference of sentences very much.

So that's all likely true. But the current discussions of exemplar-based models, they don't really discuss that very well. Like I showed you the quotations that are thrown around by a lot of people, they *don't* say, "this whole pattern-matching thing with the whole context, linguistic, extra-linguistic, social, and inferential: they are only processed to the degree that they're reduce entropy if ...". No, they don't say that. They just say, "well, everything's included, somehow rich memory representation ...", but that's not the case. So what we need to discuss is a lot of things that I've been trying to talk about this whole week, i.e. cue availability as a function of recency, association strength and things like that and then we need to quantify the power of entrenchment against recency.

To wrap up, last slide. What a lot of the work and usage-based linguistics is glossing over is exactly these things. We talk a lot about the cognitive commitment, but that means or implies that we should be developing causal and predictive models of language—just saying after the fact that something is motivated, "it fits", is not good enough. We must show that the causes we posit for something lead to the posited effects for language structure, for language use, and for language change. It's not enough to just show that the effect *can* be traced back to the posited cause.—that's not the same as showing that there is in fact a causal relationship.

## Interim conclusions

- That is, we're somewhat loosely glossing over such things when we should be
  - developing causal and predictive models of language
    - motivation isn't good enough for this
    - we must show that the posited causes lead to the posited effects (language structure, use, and change) - not just that the effect *can* be traced back to the posited cause
    - we need to ensure not only that language is predictable from the mechanisms we posit, but also that unattested results are not predicted by the same mechanisms
  - for that, we need to explore
    - the roles that information, entropy reduction, recency, contingency, salience/surprisal, & predictiveness play all together (and, sorry, with quantitative methods!)
    - accounts of memory more, and how forgetting etc. interacts with chunking & abstraction/generalization (e.g. Ellis 2002, Dąbrowska 2008)

FIGURE 33

In other words, we need to ensure, and this is maybe the most important part here, not only that language is predictable from our domain-general mechanisms, but also that things that *don't* happen are *not* predicted by the same mechanisms. Maybe hard to process: we have to show language is predictable from our domain-general mechanisms, but we must also show that the things we *don't* find in language are not *also* predicted by these mechanisms because otherwise these mechanisms are too powerful, they explain whatever the hell you want to explain with them. All the mechanisms, all the quotes that I showed you for the usage-based theory of language, they lead you to expect that backgrounds prime things, because they are part of the linguistic context. Well, but they don't. So somehow, the way things are currently being discussed is too powerful. We have to constrain it a little bit in order to make sure that we don't predict everything and their mother.

So we need to explore these kinds of things in ways that often need to be statistical and we need good accounts of memory and how it interacts with chunking and generalizations before we can be really sure that this super general model that we're building up now for the last twenty years isn't actually way too powerful, given the linguistic data we really have.

Thanks.

## Concordance, Surprisal, Entropy: Practice with R

In today's talk, we're going to look at the practice side of things that were discussed yesterday afternoon. Specifically, I want to show you a very brief and very simple example of how to compute entropy values and how to compute surprisal values. The third thing that we'll need to do, basically in order for this to work is, I want to show you briefly how I think best to do concordance output with R.

Maybe as a quick reminder: First we'll look at the concordance output. Basically, we will work on creating this kind of output. We'll be looking for some regular expression. The task will be to save the regular expression in such a way that we have a central column that contains the thing that we searched for, a column with a preceding context, and a column with the subsequent context, so that we then would be able to annotate it for whatever linguistic or contextual or other features we have in mind. That'll be the first task.

The second task will be to compute something like entropy values. What we want to do is we want to find out how diverse items are used in the slot around something that we have concordance data for. Basically, what we'll do is we'll look at the concordance, we'll look at some item in that context, and then we'll try to figure out how diverse are words used in the slot after a certain occurrence of another word.

Third, we will compute surprisal: How unexpected are certain words in a slot after another word in this particular case? Obviously, even if all the examples here are concerned with what happens after a certain word, of course you can always do that also for how often happens something within a constructional slot, within a certain textual position or anything like that. I'm just using words here because it's the simplest possible example, and because we don't have a completely ready-made parsed corpus available here for us to work with.

The first output that we want to create then is a spreadsheet like this. It looks like this. This is what we want to end up with. The main point here is the



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611591>

concordance as you see, what I'm looking for seems to be a verb, but any type of verb. We see a whole bunch of different verbs here in different inflectional forms with different tags, so this is *feels*, the tag is *vbz* for third person singular. This is *said*, so the tag is *vbd* for past tense and so on. Then as you can also see the word in front of it is always *he*. It's always *he*, tagged as a personal pronoun and so on. Actually, we're looking for two words here that you can see here at the bottom. The other word that might be in front of the verb is *she*. We're going to look at which verbs are used after the masculine personal pronoun *he* and which verbs are used after the feminine pronoun *she*. In order to make the analysis a little bit easier and to really use everything that R has to offer, on top of the data as you see them here, PRECEDING context, the MATCH and the SUBSEQUENT context, there is just the case number column that numbers all the cases from one to however many we have, about 6,500 nearly. Then there's a column here, which is just called SHE: That column just basically tells you for every row, whether it is an example, whether the verb is preceded by *she* or by *he*. These cases that you see here, all those verbs are preceded by *she*, so the SHE column always says TRUE, right? All the cases at the top, it's the opposite, so all of these are cases where the personal pronoun in front of the verb is *he*, so this column (SHE) says FALSE. The idea is that we would be very easily able to, for instance, re-sort the rows in a certain way or to only filter out the rows that have *he* in them or *she* in them. Basically what this [pointing to FALSE in SHE column] describes is what happens here [pointing to PRECEDING column]. In order for us to be able to sort, it's useful to have that in a separate column, just like any other annotation we might create. This is supposed to be the output of the first step, namely, generating this concordance.

In order to create that output, if you want to follow along in R, you will need this script file here [pointing to `og_concordance-surprisal-entropy-practice.html`]. Ideally, if you're following along like this here, you would open this again by double-clicking on it, so that RStudio immediately opens up in the correct folder. If you just follow along with the html file, that's of course perfectly fine as well.

The script that we'll be using looks like this. A few brief comments here. As usual, the first line clears memory. Just like yesterday, the second line loads a concordancing script or concordancing function from my website. Since here we want to create this concordance of *he* or *she* followed by some verb, that's something we can do with the normal functions available in R, but my function makes it more convenient, so this would be loading that function from my website, so that it's available.

Then third, here I'm defining a function that computes entropy. In base R, if you just install it from the website, there is no function in R to compute

entropy in and of itself. There are packages that contain such functions, but in fact, the computation of entropy is so simple, you can just write a function yourself. As you can see, it's just two or three lines of codes, so not really a big deal. In order for us to have it easier, I wrote this function here so that when we want to compute entropy on our input, we can just use that word *entropy* and get the results that we want.

As you will see in a moment, this function actually computes two different versions of entropy, depending on what the user says. First, it computes the normal entropy, if you will, so the standard value, and then it computes one that is normalized, which is sometimes very useful in order to compare different types of frequency distributions. We will see that here, in this particular case, this is actually important, because using one value over the other will tell you different things.

The beginning again is just as before, I've tried to make everything as comparable to what you have already seen as possible. Again, we'll be using the Brown corpus since we'll be looking for personal pronouns and verbs and things like that, obviously we're going to use the tagged version. As before, I create this vector `corpus.files`, which is the contents of the directory, `03_data`, and then `Brown-tagged`, and then `full.names` is `TRUE`. Again, we will have fifteen files to search through and all these fifteen files have the same annotation that I think, by now, probably everyone has seen. Let me show it anyway just in case.

It's again this type of annotation, as before, just to really practice it all the time, what we'll need to get rid of is the annotation here at the beginning. Then second, we will want to use the tags to find *he* or *she* and only *he* or *she* if it's followed by a verb. Essentially, we'll be describing a sequence of one word with the tag followed by—and that tag, I mean, the word should be *he* or *she*—and then there will be a tag, then there will be another word, which we don't describe, because we want every verb. We'll just use a placeholder to define that. But then whatever follows *he* or *she* needs to be tagged as a verb like here [pointing to tag `VBD`]. The one thing that all the verb tags in this corpus have in common is that they begin with an underscore (`_`), and then the next letter is a V. That's the information that will be using to find the kinds of sequences we're interested in. There are one or two other things that we need to consider again:

One is that you cannot just look for *h* and *e* followed by an underscore in order to find that personal pronoun. Why not? What's going to happen if you do that? You're going to drown in hits, because if you look for *h* and then *e* and then an underscore that will also find this [pointing to `the_AT`], right? That wouldn't work, or in other words, actually I found it very difficult to come up with other English words, but food words like *leche* or *quiche* or whatever, if they were in that corpus, they would also end in *he*. Again, we'll need to make

use of the fact that regular expressions allow us to say, find this `[h]`, and then this `[e]`, and then this `[_]`, but only if there's not another letter in front of it. The same thing as yesterday: This `[the_AT]` is not supposed to be found, because before the *h*, there is another letter indicating that this is not just the word *he*, but something else. In the same, we'll have to do with *she*. We'll again need a regular expression that says don't find this one `[the_AT]`.—I don't know whether there is a *he* here in this output, it doesn't seem like it—but only find cases when there is a space in front of the *h* and then the other tags follow.

Second, of course, we will make use of the part-of-speech tag. For searches like this, you would need to know what part of speech tag *he* takes and what part of speech tags *she* takes, in this case it's both a *PPS* for personal pronoun. Here you see an example of personal pronoun in the plural. We're going to be using this. The ultimate search expression will be something like: 'not a letter or not a number or not anything else like that in front of an *h* and an *e* and an underscore, and then *PPS* and then the space, and then some stuff that is tagged as if it's a verb'. We will not describe the shape of the verb at all, because it can be super short, like *he is*, or can be super long, *he idolizes*, I mean, as if that was super long, something like that, or *he implements*, but then whatever character sequence the verb is, it will be followed by an underscore and by a *V*. Then we'll do the same thing for *she*.

When we look at that search expression, how could we do this? How could we actually generate both these findings for *he* and for *she* at the same time? Can you imagine how you would look for *he verb* and *she verb* at the same time? It would use something we talked about yesterday briefly. Actually, I think the day before as well. Remember that we had these quantifiers as we talked about, at some point we talked about how to look for *color* both in the American and the British spelling. The difference between the two is that the *s* is optional. The difference between *he* and *she* is that, *he* is just the *she* without the *s*, so the *s* being there zero or one time: that makes the difference between *she* and *he*. Theoretically, we could look for *s*?: It may be there, or it may not be there, followed by *he* and then the underscore. We're not going to do this actually, because it makes it harder to later distinguish between matches that are *he* and the matches that are *she*, but theoretically that would be a possibility. If you really only want the concordance output, that would be a good way to go.

Let's go back to the code. This time around, we'll first create one vector that contains the whole Brown corpus. The Brown corpus is relatively small, one million words, so it's going to be no problem at all for R to hold that in memory on any relatively modern computer. That means, in this case, we're actually not even going to search for something like we're not going to extract words, or we're not going to do much of looking things up or anything like that—we're

just going to load every file, clean it by removing the line initial annotation, and then save it for later.

Whenever I say something like *I'll save it for later*, we have multiple files, we again need to create a collector or a container structure that will take up all the words, all the sentences from the file. That's what I'm doing here. `all.corpus.sentences` is supposed to be `[<-]`, and then the `c` means 'combine' or 'concatenate' or something like that. But in this case, we're actually not concatenating anything. At this point, this is an empty vector that is ready to take up stuff from whatever processing happens at a later point in time.

Then the first step is the same as always. We have a loop, `for`, and then a counter that goes through every one of those fifteen files, so `1:15`. As before the whole time we load into an object called `current.corpus.file`, we load `corpus.files[counter]`, the first, the second and the third, then the fourth version, which is a character string separator as a line break and `quiet=TRUE`. No problem there.

The next thing, again, will be that we remove the line-initial annotation. Remember when we looked at this last time, we've already solved this problem before. The line-initial annotation we don't want is everything from the beginning of the line to the first space: beginning of the line to the first space. We can simply recycle the code from before and say: the contents of the current corpus file is now supposed to become what happens when you replace this by nothing, in `current.corpus.file`, `perl=TRUE`. The old version of `current.corpus.file` gets overwritten by a new one that involves this replacement and again, the regular expression is: the caret `^` means 'at the beginning of the string', the period `.` means anything, then the asterisk `*` means 'zero or more times', and the question mark `?` means 'till the first whatever follows next'. A different way to read it would be: the sort of things at the beginning of the line until you reach the first space, all that will be replaced by nothing. We are getting rid of the line-initial annotation. Then we just collect the contents of the current corpus file in this vector, `all.corpus.sentences`, that at the end of the process will contain all the lines of the Brown corpus.

If we run this here, already finished. How many lines does it have? We can very quickly check this. We have 57,000 lines available here. This is the beginning—it's already too long actually to show on that screen right now—this is the beginning of the corpus. It looks fine, and it seems like here, in this format, every line is a sentence. It always begins with something. Then, there's a period at the end, period at the end, period at the end. It seems like the loading of stuff was in fact successful. Now, if we look at this, just to maybe remind you: can you imagine how you can very quickly find out how many words we have here? What would you need to look for to find out how many words we

have in this corpus? As an approximation at least. One easier approximation would be to just count the number of underscores, right? Because every underscore is what separates a word from the tag. That will slightly overestimate the number of words, because punctuation marks also have their own tags. But, for instance, if you were to look for sort of any letter in front of an underscore, that would give you the number of words of the Brown corpus very easily. That would be another simple way to get at what a frequency list might tell you.

Now we want to do something else: We want to find the verbs after *he* and the verbs after *she*. That's what we're looking at here. Let me first show you how you would do that in R in the normal way before I then show you how we do it in a better way.

The normal way would be something like this. Let's say `verbs.after.he` is something like and then the normal looking-for-stuff function in R is called `grep`. We would look for something in `all.corpus.sentences`. Actually, I'm not sure I ran this one. What would one be looking for? Typically, it would be looking for something like this, *he* followed by its tag, followed by a space [`"he_pps "`]. Then we would have to describe the word, which is followed by its tag which begins with `v [_v]`. How do we define a word in this corpus? How would we find the verb, which then has its tag? Whenever we look for something here, we always have to consider that it consists of two parts. Every word will have its tag. If we look for *he* and some verb, we will need to define *he*, which we're doing here, and then its tag, which we're doing here. Now we need to do the same for the verb: We need to define the verb and then the tag. The tag I've already defined, it's an underscore and then `v` for verb. But how do we tell R to find the verb itself?

One smart way to work with corpus data like this is always to bear in mind, what is annotation. Annotation is separated here—I mean the annotation is the tag—and it is separated from the word in question with the underscore. So that when you read it, you know that *he* is the word in question, and this `[_]` is the separator, and this `[pps]` is the tag. That's what the underscore is for, to introduce a character between the word and its tag. But what that means is that every word will be 'stuff that is not an underscore'. Because the underscore in this corpus is used to separate words from tags. So the underscore is a character that, you know if you see it, that's not a word. In other words, if you don't see it, that will be part of a word. What we just need to say is something like this `[_]+` a character class, so any kind of character, that is not an underscore, one or more times.

Let me show you how this works in this other application that I showed you the other day. If we have a search expression like this: `he_pps [_]+_v`. And then we have text. [...] This finds it. Again, in corpora like this, that have



an annotation, where the annotation follows or precedes the word, it's always safest to define words as 'not being annotation'. So this *he* here matches that *he*, `_pps` matches this and now we're skipping over the *said*, the `_v` here matches this `_v`, and the *said* the *s*, the *a* and the *i* and the *d*, those are matched by 'we're not underscores'. That's how you would look for things like that. The same applies to, very usefully, to corpora that I'm using SGML/XML annotation. So where the parts of speech, for instance, are between angular brackets in front of a word. There you can always say that a word is anything that is not the next opening angular bracket. Because the next opening angular bracket will be part of the tag for the next word. So defining words as 'whatever is not annotation in this corpus' is usually the safest route to go. Again, typically you would do it in a way like this. Now we already have it. Let's see how many such cases there are, verbs after *he*, `length(verbs.after.he)`, so nearly six thousand occurrences of verbs after *he* like this.

Let's look at the first few. For instance, here's the first match, right? "`he_pps added_vbd`". But so this would be the normal way to find this in R, but this is of course very annoying, because now we get the whole sentence, but we do not get *he* and the verb sort of specially separated from the rest. Usually, what we want to see is we want to see the item that we're looking for in this middle column and then the preceding context and then the subsequent context. Here we don't see it like this: R just throws the whole sentence back at us without any additional structure that we can use for sorting or anything like that.

That's why we *don't* want to do it this way because it would make, I mean we can save this very easily, but it's very hard to process afterwards. For instance, you can see this line, for instance, is quite long. If you want to annotate this, you have to read the whole long line to even find where the *he* is with a verb afterwards. That obviously doesn't make for a nice, effortless annotation.

We're going to do something else. That's something else involves a slightly more complex, regular expression, which looks like this. Let me show you this in here. We're at this point right now, this is the thing we will be talking about in a moment [ "`(\"?<=\\bhe_pps \")[^_]+(?=_v)\"` ]. I'm defining an object, `verbs.after.he`, and I'm using this function of mine `exact.matches.2`. We're looking for something—I'll explain this in a moment—and we're looking for it in `all.corpus.sentences`. Then actually we do the same for verbs after *she*. We're again looking for something in `all.corpus.sentences`.

What does the regular expression? What is highlighted here at the top? This kind of stuff, you first look at this to get an idea of the structure of the regular expression. If you look at this regex, and you know a little bit about regular expressions, you see that it has three parts. The first part is this parenthesized

unit, the second part is the stuff in the middle that is not parenthesized, and the third one is the part of the end that is parenthesized. Those are the three parts you want to interpret.

This one `[^_]+` is really easy because we actually just talked about it. This `[^_]+` means ‘not an underscore one or more times’. But then there’s all the other stuff, so the stuff in front of it and the stuff after it. Those things are what are called instances of *lookaround*. So *lookaround* is a more advanced regular expression concept, a construct that allows you to say, ‘I want to find something, but only if it’s located in the vicinity of other things that I’m describing’.

What this means here is ‘find things that are not underscores if you can look to the right, but only if, when you have them, you can look to the right and see this, `_v`’. So you only want to find a word if, on the right of the word, there’s a verb tag, that’s what that means. The *lookaround* expression here is the parentheses, the `?` and the `=`. That says from where you are right now as a regular expression engine, when you can look to the right of that, then you must see the `_` as a tag indicator and the `v` for I’m a verb.

At the same time, we also are using *lookaround* in front of this expression. That’s all this long thing here, but it’s actually relatively simple. It’s the opposite of this. It’s when you have found something look to the left—see here there’s the parenthesis, just like here, and then the `?`, just like here, then instead of an `=`, there’s an `=` with a `<` in front of it. The `<` points to the left, so it says to the left of me, there has to be something. What is supposed to be to the left of me? Well, the *he* followed by `pps` and then remember this, the `\\b` says *he* must not be part of a word, there can’t be another letter in front of it, or a number in front of it, there has to be a word boundary. Essentially, what this means is, I’m going to try and show this here with highlighting. This one, so try to follow along with the highlighting, with the mouse marking here. This `[^_]+` says find something that’s not an underscore. This stuff `[added]` is not an underscore, but now only if you can look to the right and see `_v`, and yes you can do that. If you highlight this `[added]`, then you can look to the right and see `_v`, meaning that this `[[referring to “added”]]` is a verb, but then also only find this if you look to the left and in front of it, in the reverse order, you find a space, a `s`, a `p` and `p`, then `_`, then `e`, then `h` and then a word boundary. That’s how we’re finding this instance here. Is that clear?

Since in English it’s so nice, it’s so similar. If we want to do the same thing for *she*, we just recycle the whole thing and put an `s` in front of it. Then we’ll find *she* plus verb. *Lookaround* is a little bit more advanced as a regular expression, but it allows you to really, as you can see, customize things very nicely. We’re getting exactly that one thing in the verb slot here `[added]`, depending

on what's following [\_VBD] and depending on what is proceeding [SPP\_eh] the expression in question. If we run this, nothing seems to happen, but we now have two objects. One is called `verbs.after.he`, one is called `verbs.after.she`.

Now we do have all the contents that we're interested in, but we have them in a format that is not yet ideal. Let me show you what we have in fact. There's a very useful function in R called `str`, which gives you the structure of an object. What is the internal structure of an object that you created? The function that I wrote, that we used here, `exact.matches.2` returns a list that contains multiple items. The first item is the verbs that we looked for, *added*, *resigned*, *scheduled*, *told*, and so on. The second item is where in the input are these lines located, there's one in line 75, one in line 88, one in line 90, and so on. Then it gives you a very small or very crude dispersion measure, which I'm going to ignore for now.

Then this is the most important part right now, it gives you this component. The fourth one is called 'lines with delimited matches'. What that gives you is exactly this output here, the one that we're trying to get in the spreadsheet software like this. For every single time that there is one of these personal pronouns followed by a verb, what this fourth component contains is essentially this: PRECEDING context, the MATCH, the SUBSEQUENT context. It already gives you that with the tab stop, so that you have an easy way to put this into a spreadsheet software and have the verb be separated from the preceding context and the subsequent context. Let's see whether we can see that here in the output actually. Yes, it's even easier to see. The first match is, *but* and then *he added that none of georgia's congressmen....* That's the context. So see what happened here: R, the function, took the *he* and its tag, and then it inserted a tab stop in front of and after the verb, so that we get this output here, *he* and then the column break, the transition from one column to the next, that's the tab stop, then we have the MATCH, then we have another tab stop and another column, and then we have the SUBSEQUENT context.

What we're doing now in the script is, we're creating a concordance element, `verbs.after.he.she.concordance`. We're just combining `verbs.after.he[[4]]`, the concordance output, which is in the fourth component and `verbs.after.she[[4]]`, the concordance output, which is in the fourth component. Let me show you this. If we run this again, nothing seems to happen. It just does it. But now let's look at the first two elements, for instance, first two elements, one and two. It looks like this. So now we have the tab stop here, which will be a column. Same thing here. Here is the *he* and it's followed by a tab stop, the `\t` is a tab stop, *resigned*, that's the verb. Then again, there's a tab stop, and then there's the rest of the context. The function basically takes every

sentence and inserts the tab stops around the things you've been looking for, so that you can put it in a spreadsheet relatively easily. I'll show you how that works in a moment.

Then the final thing we want to do is we want to add the column that you've seen here [SHE column]. We want to have a nice and easy identifier for every instance, whether it is one of the *he* matches or one of the *she* matches. This is how we do this: `verbs.after.he.she.concordance` becomes the result of, now you're seeing a new function here, the function is called `paste`. What `paste` does is it takes separate character strings and glues them together into a longer character string. It might take two short things, you have one element here, another element here, and it puts them together.

Let me show you. You have one word and another word [`paste("he", "stinks", collapse=" ")`], so now it's one string ["he stinks"]. Here it's two strings, one ["he"] and two ["stinks"]. Now I say collapse those two, put a space in between them, so now it's one string ["he stinks"]. That's what we're doing here. So what am I gluing together? I'm gluing two things together. First, the concordance, so PRECEDING context, MATCH, SUBSEQUENT context. I do that for all like seven thousand or six thousand or however many it was together at the same time. Then I put it together with something else, namely, I look for, in the concordance, whether there is *she*, personal pronoun, and a tab stop [`"\\bshe_pps\\t"`], and I look for that in the concordance. Basically, what I'm doing is I take the concordance, and now I check for every line, whether it's a case of *she* or whether it's a case of *he*. I'm doing that with the `grep1` function, because that says if there is a *she*, it will say TRUE. If there is not a *she* like that, it will say FALSE. That will basically disambiguate every line by saying, this is a line that was generated because of the *he* in there, this is a line that was generated because there's a *she* in there and so on. If we do this again, let me show you what that does. We're putting it together like this. Again, the first two instances now look like this. We've added the FALSE here at the end. It says FALSE, because this whole line doesn't have a *she* followed by a tab stop, it has a *he* in there. So it's not a case of the feminine pronoun, it's a case of the masculine pronoun. Same here: This whole line does not have the *she* plus personal pronoun tag in there, but *he*, so this says FALSE. If we now print that into a file, that separate column will tell us what kind of concordance match it is. This is how we do this. If you run this script, you will get the same output file.

The function to print something into a file in R is `cat`. What I'm printing now here is first I print a column header, "`PRECEDING\\tMATCH\\tSUBSEQUENT\\tSHE`". That's this here. That's what I'm printing first. After that, I print the whole concordance, all 6,500 items, everyone of them is separated from the others by a line break. I print it into this file

[09\_concordanc-surprisal-entropy-practice.csv], 09 because this is the ninth talk, concordance surprisal entropy practice csv. If you do that, let me show you what kind of file you'll get.

So that file looks like this, also known as 'terrible'. It doesn't seem like this is particularly useful, but look: the point here is to recognize that you can see that here a little bit. Here's a *he* and then you see here's this arrow →, that's not a space, so that's the tab stop. Then here's the verb [said], and there's the other tab stop [→], and then here is the verb tag [\_vbd]. So we can already see some structure in there. It's just that in a text editor, it doesn't come out very nicely. What we're going to do is this: we're not going to open it in a text editor. Let me see what happens. If you have a file csv and that's why I gave it this ending, then most of you are probably using Excel or something like that. If you have that installed, then double clicking on this file will open a spreadsheet software, specifically an assistant to enter this file into a spreadsheet software.

Look at this, double-click on this, and then see what happens here: This text import assistant opens up, and your spreadsheet software will try to guess what is the structure of the file that you are trying to open. You already can see a preview down here. Now you have to declare to your spreadsheet software what is the structure of this file. This one is important, because you don't want to make any mistakes here. In this case, LibreOffice guesses, and it guesses incorrectly, actually, it guesses that the things that I have between columns are tabs—that is right, I put those there—but it also guesses commas, which of course is not right. I mean these are corpus files, the commas are actually just commas from the text. You want to deactivate that and you see here it makes a difference. For instance, here *the bill*, and then there's nothing. There isn't a *he* at the end so that's kind of wrong. But if we say 'don't split on commas', now you see the remaining context and now the preceding context actually does end in *he* with a personal pronoun tag.

If you open a file like that, you want to make sure that only the tab button here is ticked, so that the spreadsheet software recognizes 'only the tab stops are really the delimiters here'. Then you click OK, and then open it here in the other window, and then it looks like this. That's still pretty ugly, but it actually is the right format, because now look with just a few clicks, you can make it look nice. This column, as you can see, is super wide, because there are some really, really long contexts. The first thing you're going to do is you're going to make this narrower, so not forty inches, no, but three. Here's the other one, we make that just as wide, also, let's say just three. Then we center this [MATCH column]. We right-align this [PRECEDING column], and now it already looks much better. Now, you can read it nicely, especially if we make it a little bit bigger like this. *In two other cases, he ruled that ...*, now it looks nice. You can do

other things like make this bold and all sorts of other things to get the format right. This would be the way that we get a concordance into R: You basically create the concordance with this `exact.matches.2` function. Then, from that output you extract the fourth component, and then you print that into a file. As soon as you double-click on that file, in your case, probably Excel, you will open this text import assistant and you just need to say, whatever you're trying to open right now, it is split up by tabs and then you'll get this nice output. That will be the way to go. Let's get rid of this.

The next thing we want to do is we want to check which of the verbs co-occur most with *he* and which of them co-occur most with *she*. Of course also, how diverse are these verbs after the two personal pronouns.

The first thing we'll do is we create a vector called `verbs.after.he`, which takes from the concordance output only the verb, then we do the same for `verbs.after.she`. From the concordance output, we only take the first element which is what contains the words. What do the two vectors now look like? Let me show you.

Now we have a vector `verbs.after.she` that contains every verb that has ever been used after *she*, same for *he*: *added*, *resigned*, *scheduled*, *told* and so on. What we want is a table that has all the verbs in the rows, then *he* and *she* in the columns, then we see the co-occurrence frequency of the two. Again, the nice thing is you can use what you've already learned in the previous sessions. We're adopting the same logic here as before, namely, doing this. What does that do? First, we take all the verbs after *he* and put behind them all the verbs after *she*. We have one long vector with all the verbs used after each of the personal pronouns. Then, we create a second vector that is just as long. That says, for every verb where that has been used with a *he* or with a *she*. The second vector will say *he* a few thousand times, then it will say *she* a few thousand times, namely indicating for every verb what it goes together with in the concordance output. That's what happens here: `all.verbs` is the combination of `verbs.after.he` and `verbs.after.she`, done, you take all the verbs after *he* and add the ones after *she*. `all.pronouns` is, see, we've done this before last time, repeat the words *he* and *she* as many times as there were verbs after *he* and as there were verbs after *she*.

Now we're creating this 6,500 items long vector that says *he* 5000 times and then *she* 1500 times or something like that. Because then we can tabulate and see `verb.by.gender` is a table that cross-tabulates all the verbs and all the pronouns. Let me show you that.

This is what we want. We now have a table that tells you, the word *said* occurs this many times after *he*, this many times after *she*; the word *thought* occurs this many times after *he*, this many times after *she*. As you can see here, actually it seems as if the pronoun *he* is much, much more frequent. Every

one of these words, if I see it correctly quickly now, has much, much higher occurrences of *he* than of *she*. The reason is this: There's approximately five thousand cases of *he* in the corpus, but only fifteen hundred cases of *she*. *He* is about 3.5 times as frequent as *she*. That's why we have this huge discrepancy here in every one of these cases. It's not like *went* or *saw* is actually so much more typical of *he*. I mean it said *he* is just used more often in general. We can actually not just take this frequency and make any conclusions on the basis of that—we would have to do something better than this.

Now if we have a table like this, we can actually already easily compute something like entropy. For instance, we have this long vector `verbs.after.he`. We just create a frequency table which will lead to these frequencies. Every verb, how often is it attested after *he*? Then we can use the function `entropy` that I gave you at the beginning. This value here says eight point five [8.493393]. That doesn't actually mean much—it's not like that's significant or not significant, or important or not important. It only becomes interesting when we compare it to the corresponding value for *she*. That value is lower [7.637634]. The entropy after *he* is higher than the entropy after *she*, which, given what we talked about yesterday, for instance, would mean that if you hear *he*, it will be harder for you to predict the next verb than if you hear *she*, that's what that says. There is more diversity, there's more uncertainty in the slot after *he* than there's after *she*. Again, remember, I'm using *he* and *she* here and the slots after, but of course this would also apply to constructional slots, if you look at a syntactic construction or something like that.

For instance, to use what Arie talked about yesterday, grammaticalization, to the extent that something becomes grammaticalized and sort of loses semantic specificity, you would expect that to correspond with an increase in the entropy of that constructional slot. If something gets very general semantics, then that means it can co-occur with many more different things. That would correspond to an increase in entropy. I mean that would at least be the prediction.

So there's more diversity, more uncertainty after *he* than after *she*, but this is actually not yet the right analysis. That is because the entropy value, the way it is computed here, is related to the overall frequency. We've seen that there are more than three times as many *hes* as there are *shes*. So part of that difference might actually just be a reflection of the fact that *he* is so much more frequent than *she*. The way to correct for this is to use this other version of entropy that my function can help you compute. We actually do the same thing again, but now we say `entropy of this frequency table normalized=TRUE`. So what the normalization does is, it brings entropy down to a normalized range from 0 to 1, and one that is not correlated with the frequency anymore. Now you can actually see that the result now turns out to be the opposite. Now the

uncertainty is higher after *she* [0.8566328]: Once you correct for the fact that *she* is rarer in the corpus than *he*, then actually the diversity after *she* in terms of verbs is higher than after *he*. Another good reason for what I said earlier, don't use, I mean what I've been saying the whole week, 'don't use measures that conflate frequency with other things'. I mean this way of computing entropy conflates frequency and diversity. If you want to make sure that you get really only diversity information, then compute the normalized version like here, and you get a better picture of what actually happens in that particular syntactic or constructional slot. So the verb used in this corpus after *she* is more diverse than the verb used after *he*. But of course that does still not tell us what are those verbs. Which verbs seem to be somewhat characteristic, at least of *he*, and which would seem to be somewhat characteristic of *she*? This is now where we might look at surprisal.

Like I told you yesterday, surprisal is actually really easy to compute. You just take the log of a percentage, I mean, the binary log of a percentage, and make it negative. What we're going to do is we're going to start out with this table `verb.by.gender`, which is this table. That thing is called `verb.by.gender`. It has a column called *he*, and it has a column called *she*, with all those frequencies here. So if we want to compute surprisal, we first need to convert these into percentages. How do we do that? Well, we have [4,968] cases of *he*. So every one of these numbers needs to be divided by [4,968]. So then this will be what like eight percent or seven percent or something like this. This will be much less, and obviously all these here will be even less than that. This will be like one percent and so on.

Then we just need to log that. If we do that, it's actually really simple. We're now saying, the surprisals for all the verbs after *he* are going to be the negative binary log of the frequencies of the verbs after *he* divided by the overall frequency of *he*. This is doing the division by the total. Here we do the same thing. The surprisal values for *she* are the negative binary log of the frequencies of verbs after *she* divided by all the instances of *she*. Then we got it.

If we then sort it and look at the top thirty, we can see which verbs are most surprising after *he* and which verbs are most surprising after *she*. For some weird reason, the word *married* is really surprising after *he*, but not after *she*, I wonder how that is possible. But so these verbs here are really not expected at all after *he*: *clung*, *described*, *developed*, *married*, *settled*, *sarled*, whatever.

Then, these are funnily enough, the most surprising verbs after *she*. For some reason, *explained* is really surprising in the Brown corpus to find after *she*, that's not something that happens there a lot. Then same with *pulled* and *sees* and *considered*. Now if you were into gender studies or sociolinguistics or something, you can try explore all those and construct a story what that



shows you about maybe, gender stereotypes in American English in the 1960s or something, if you're interested in that.

Now these are all, as you can see, these [surprisal=10.53722] are actually all the same. The reason for that is that those are probably verbs that all show up just a single time. It's always one divided by the number of times that *she* occurs, so it's not particularly discriminatory here because we have so many hapaxes, so many cases that occur only maybe a single time, or maybe two times or something like that. Obviously, for a more detailed analysis, you also would want to scroll down a little bit further to see which of those verbs are frequent as well. Of course, ideally what you would do is look at the surprisal values and frequency at the same time.

So how would you do this? Let me show you one possibility at least very quickly. So we have the frequency of the verbs after *he*, and we have their surprisal values. If that doesn't scream out for a two-dimensional plot, then what does? What we could do is plot verb by gender, the values from the column *he*, and let's log them, and surprisals *he*. You can see, actually, that in this case, there's a perfect correlation between the frequency on the one hand and how surprising they are. In this case, that's of course because the surprisal values are based on the frequency, namely, I mean the negative log of that. So it would be actually more interesting to compare this with the overall frequency. Let's look at that: We find that some of the less frequent words differ quite a bit in terms of how surprising they are. For instance, for words that occur four times or between four and eight times, they can be quite differently surprising. Now the most interesting thing, of course, would be if we actually plot the words there. We can see which words are super surprising, and which words within a certain frequency range occur with that frequency, but are actually more surprising or less surprising than other verbs.

If we did this for both genders, both *he* and both *she*, then of course we might actually end up with some implications of this, if we are interested in stereotypes like that.

Again, just to sum up, basically, so relatively simple. If you want to compute entropies: as soon as you have any kind of frequency vector in R, you can just run that function that I gave you on that. If you have any kind of percentage values, you can immediately log them to the base of two, make them negative, and you have these surprisal values.

Like I talked about yesterday afternoon, this will give you some idea of things, especially if you apply it to constructional slots, not just to words of some of the things that are deeply relevant to matters of processing, matters of acquisition, and maybe also matters of language change. Let's leave it at that, and let me know if you have any questions.

## Corpus-Linguistic Applications in Cognitive/ Usage-Based Explorations of Learner Language

Thank you very much for the introduction and thanks for all of you making it to the grand finale.

In this talk, basically, what I want to do is I want to show you a few case studies in which many of the things that I've been talking about here are brought together.

Essentially, I want to show you examples of how contingency, recency, all these kinds of things, surprisal, can be integrated into actual analyses of linguistic phenomena. Many of my talks before have been primarily methodological, always with an attempt to inform cognitive linguistics or psycholinguistics

**Overview of this talk:  
bringing many things together**

- This talk is about bringing many, though not all, of my above admonitions, requests, & wishes together ...
- Three case studies, all involving learner English
  - *that-complementation: native vs non-native speakers*
    - comprehensive annotation of concordancing
    - involves association ( $\Delta P$ ) & surprisal
    - more advanced statistical approach
      - imputation & mixed-effects modeling
  - *genitive alternation: native vs non-native speakers*
    - comprehensive annotation of concordancing
    - more advanced statistical approach
      - imputation & mixed-effects modeling, →
      - individual variation
  - *priming in learners' dative alternation*
    - corpus data from native speakers and its relation to native-speaker AMS ( $\Delta P$ )
    - more advanced statistical approach
      - mixed-effects modeling
      - within-experiment learning (recall recency?)

FIGURE 1



All original audio-recordings and other supplementary material, such as any hand-outs and powerpoint presentations for the lecture series, have been made available online and are referenced via unique DOI numbers on the website [www.figshare.com](http://www.figshare.com). They may be accessed via this QR code and the following dynamic link: <https://doi.org/10.6084/mg.figshare.9611630>

or bring these things into corpus-linguistic approaches but so this time we're actually going to look at a bunch of studies and the data and the kind of things that you can do if you involve the methods that I've been talking about here in your research.

There's going to be three case studies, all involving learner English so all based on, to some extent, at least corpus data involving learner English. Two studies are actually also about experimental data. I'll talk a little bit about how the corpus data studied here intersect, or can be brought together, with experimental things. The first case study is going to be on *that*-complementation. The question there is going to be looking at the behavioral differences or the different linguistic choices made by native and non-native speakers.

What that example will feature is, for instance, the comprehensive annotation of concordancing of many thousands of examples of concordance data, actually. It will involve association measures, in particular  $\Delta P$ , it will involve the notion of surprisal, and it will involve a relatively advanced statistical approach, namely, mixed-effects modeling on the one hand and, secondly, a newly developed method that basically involves something like missing data imputation. That will become clear in a second.

The second case study will be on the genitive alternation. Again, the contrast to be looked at will be the one between native and non-native speakers. This one, too, involves, relatively comprehensive annotation of thousands of examples from corpus data. It, too, involves the same kind of statistical method, namely imputation and mixed-effects modeling. But it will also give us a very brief glance, at least, at individual variation of the type that Dąbrowska, for instance, said cognitive linguistics should be more involved with.

And then the final case study will be on priming in learners of English as dative alternation. Here what we're going to look at is corpus data from native speakers, and then the relationship to the native speaker data and the non-native speaker data also in corpus data. At the same time, we will look, or correlate the corpus data with experimental data. The experimental analysis involves mixed-effects modeling as well as a statistical control for something that I've told you about, I think, twice, namely this idea that, within-experiments, already there might be learning going on. So, what I'll show here a little bit is how one can try to tame this kind of variation in a statistical fashion.

The first case study is that of *that*-complementation, as I mentioned. Basically what we did is we look at three different kinds of complement constructions.

The most frequent one was object complementation, so the alternation between *I thought* and then *that* or not *that Nick likes candy*. The question is,

## The phenomenon in L1 English

- We are exploring the factors that govern the variable presence of the complementizer *that* in English (Wulff, Gries, & Lester to appear)
- object complement constructions
  - I thought *that* Nick likes candy
  - I thought        Nick likes candy
- subject complement constructions
  - The problem is *that* Nick doesn't like candy
  - The problem is        Nick doesn't like candy
- adjectival complement constructions
  - I'm glad *that* Stefan likes candy
  - I'm glad        Stefan likes candy
- many studies have explored this variation for NS (Thompson & Mulac 1991, Tagliamonte & Smith 2005; Torres Cacoullos & Walker 2009; Jaeger 2010) ...
- ... but there are very few for NNS (but see Durham 2011; Wulff, Lester, & Martinez-Garcia 2014, Wulff to appear)

FIGURE 2

what are the factors that determine whether a speaker inserts *that* there or not? Second one would be subject complement constructions so *the problem is*, and then *that* or not *that Nick doesn't like candy*. Then the final one is actually extremely rare in our data: *I'm glad that Stefan likes candy*, or *I'm glad Stefan likes candy*, with or without *that*.

So all the time, the question is, what is it that makes people insert the complementizer or leave it out? There's a lot of studies on this that look at this for native speaker data, but there's actually relatively little in terms of non-native speaker variability. There are some studies, but not many.

If we look at the factors that affect whether the complementizer *that* gets realized or not, there's a bunch of factors that have been discussed in the past. I'm going to use this highly artificial sentence to show you how some of these factors operate. Before we get to the concrete example, one thing we can already see is that, in writing, the proportion of *that* is much higher than in speaking and in formal language, across speaking and writing, the proportion of *that* is much higher than in informal language. But apart from this general factor, there's also a lot of factors that have to do with the specific linguistic utterances that have been produced, not the general context of utterance production.

So one is, for instance, the complexity or the weight or the length, whatever you want to call it, of the subject, both of the main clause—so the *I*, right?—but also the subject of the complement clause, in this case, the *he*. The general tendency that you find is that the heavier or longer or more complex these

### Complementizer realization in L1 English

- Many variables have been shown to affect / be correlated with *that*-realization in L1 English
- *Seriously, I really hope very much that he likes it*
  - **MODE/REGISTER**: writing ><sub>that</sub> spoken / formal ><sub>that</sub> informal
  - **SUBJECT<sub>MC/CC</sub> COMPLEXITY/WEIGHT**: heavy ><sub>that</sub> light (esp. *I*)
  - **CLAUSE JUNCTURE**: intervening material ><sub>that</sub> none
    - e.g. between matrix clause subject and verb
    - e.g. between matrix clause verb and the *that* slot
    - e.g. between *that* slot and the complement clause
  - **CLAUSE-INITIAL MATERIAL**: yes ><sub>that</sub> no
  - **MAIN-CLAUSE VERB**: infrequent ><sub>that</sub> frequent (truth claims)
  - **SURPRISAL**: some verbs 'prefer' complementation patterns and 'anticipate' continuations
  - **INDIVIDUAL VARIATION**: some speakers prefer using *that*

FIGURE 3

subjects are, the higher the probability that speakers insert *that*. In this case, actually, *I seriously hope that he likes it*, the probability of *that* is relatively low, because *I* and *he* are so short and simple. Right?

Then, the question of clause juncture: Is there intervening material between, for instance, the subject of the matrix clause and the verb of the matrix clause? In this case, there is an adverb intervening between the subject and that verb. If there is any intervening material, *that* becomes more likely.

The same thing for intervening material between the verb of the main clause and the *that* slot. In this case, there's *very much* between the verb of the main clause and *that*. Again, the more material there is, the higher the probability speakers will insert *that* there.

Finally, intervening material between the *that* and the subject. So in this case, there's nothing, right? Again, the more material there is, usually the higher the likelihood that *that* is inserted. This is already a pretty good case to show why you do need statistical analysis for something like this: Even in this simple sentence and with only the few factors that we looked at, already some of the factors clearly vote for a *that* to be inserted, others clearly vote against that. In order to find out what is being done here now, I mean, on the whole, some sort of statistical analysis will be necessary.

Then there's the question of clause-initial material. Is there anything preceding the main clause even? If there is something, yes, then the probability of *that* is higher than if there isn't something. Then, there is an effect of the main-clause verb: If it's less frequent, people put *that* in more often. If it's more frequent, then they don't. Then, also some verbs prefer a certain complementation

### Complementizer realization in L1/L2-German Spanish

- In L1 German, *dass* ('that') is optional in S/DO complements, but obligatory in ADJ complements
- in L1 Spanish, *que* ('that') is obligatory in all three contexts
- in L2 English,
  - Durham (2011) compares L1 English and L2 English by French, Italian, & German learners
    - FR/IT learners ><sub>that</sub> GE learners & NS
    - other main-clause subjects and verbs ><sub>that</sub> *I think/hope ...*
    - GE/IT learners ><sub>sensitive to CLAUSE JUNCTURE</sub> FR learners
  - Wulff, Lester, & Mart.-Garcia (2014) compare written L1 English and L2 English by Spanish & GE learners
    - intermed.-adv. learners use *that* similarly to NS but more and are more impacted by processing-related factors
  - Wulff (to appear) adds spoken data to the previous study and arrives at similar conclusions

FIGURE 4

patterns and anticipate continuations, which is where surprisal comes in. The degree to which you can, once you see the verb, already guess that there will be a complement clause, that has an impact on the *that*-realization as well. Then of course as always, there is an individual variation. Right?

Again, in this one example, we already see some factors that say put a *that* in, right? There's intervening material here and here. On the other hand, we have factors that say don't put *that* in. There's nothing here and the subjects are short so the task will be to figure out what would happen in a case like this.

If we look at this in L1/L2-German Spanish, then we find that in German, actually the *that*, the corresponding element is optional in subject and direct object (S/DO) complements, but it's obligatory in adjectival (ADJ) complements.

In Spanish, *que* is obligatory in all three contexts. So that of course might be something that we would expect to maybe show up in the Spanish learners' use of English. Because if, in their L1, the *that* element is always obligatory, then maybe they will overuse it even when they speak a language in which it's often optional, like in English.

Now in L2 English, the few studies that we've had, there's one in 2011 that compares L1 and L2 English by a bunch of different learners. She [Durham 2011] finds general patterns of overuse and underuse: French and Italian learners use *that* more often than German learners and native speakers. She [Durham 2011] finds an effect of main clause subjects and some sort of verbs, and finds that, especially with *I think*, and *I hope*, *that* is really rare. Then, she finds that German and Italian learners are sensitive to clause juncture, so to interruptions or intervening material, more so than French learners, but the

study didn't really have a particularly fine-grained resolution on all the different factors that have been attested.

In another study [Wulff, Lester, & Mart.-Garcia 2014], the written L1 English and L2 English by Spanish and German learners was compared, in a fashion that's actually somewhat similar to what we have done. They find that intermediate to advanced learners use *that* similarly to native speakers. On the whole, the proficiency is relatively high, but they are more impacted by processing-related factors: So if speakers don't speak in their L1, then they speak probably under a higher processing load, because they have to formulate everything in not their first language so any factors that have to do with processing costs then might exert a particular toll on the speaker and lead to non-native speech.

Then, another study [Wulff, forthcoming] added spoken data to the previous one, to this one here that was only on written data, but essentially arrived at similar conclusions.

Now what we want to do is, first, we want to answer the question, what factors govern *that*-realization. But then secondly, in terms of methods, even more importantly, when and how do learner choices differ from those of the native speakers? We're trying to be better than the previous study by including surprisal as a predictor, which of course is particularly relevant in this context, because I've been harping on this the whole week. We're using a better statistical method: instead of a binary logistic regression that ignores the repeated measurements by different learners, we're using a two-step regression method called MuPDAR that I will explain a little bit later, and we're using multi-level or mixed-effects modeling for both regressions to take into consideration

### The present study (compared to previous ones)

- Here, we are addressing the following questions
  - what factors govern *that*-realization in intermediate-level GE and SP learners?
  - when/how do learner choices differ from those of the NS?
- we are trying to improve on Wulff (to appear) in several ways
  - we include **SURPRISAL** as a psycholinguistic predictor
  - we are using a more sophisticated statistical method
    - instead of a traditional binary logistic regression, ...
    - ... we are using
      - a **two-step regression procedure called MuPDAR** (see below)
      - **multi-level/mixed-effects modeling** for both regressions (which allows us to take into consideration variation at the levels of speaker/file, verb token, and verb lemma)

FIGURE 5

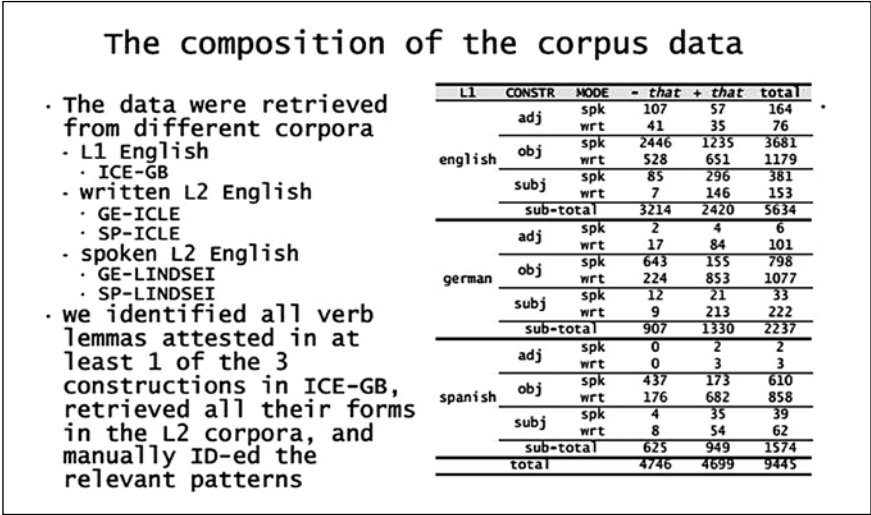


FIGURE 6

variation on the level of the speaker, which will here be the file, but then also verb tokens and verb lemmas.

So what are the corpus data? We looked at different corpora, so for L1 English, so native speaker comparison data, we took the corpus that you've seen mentioned multiple times throughout this week, the British Component of the International Corpus of English. For written L2 English, we took the German and Spanish Components of the ICLE, that's the International Corpus of Learner English. Those are written essays composed by learners of, in this case, German and Spanish L1s. Then, we had spoken L2 English from the LINDSEI corpora, again for the German and for the Spanish learners. What we did is we used the parsed annotation for the British Component of the International Corpus of English and looked what are all the verb lemmas that in that corpus are attested with that complementation at least once. That's a kind of method that I talked about, I think yesterday, where basically we don't have an annotated learner corpus for complementation. Obviously, we cannot find cases where *that* is not realized by looking for *that*, because we're not going to get that so we took the verbs that native speakers use with *that* complementation and looked for those in the learner data, and then read through all the examples to find out whether the patterns that were returned were in fact complementation constructions of the three types.

This is the detailed breakdown of the corpus data. The two main points to consider here is first, somewhat amazingly, the two constructions without *that* and with *that* are nearly equally frequent. That's rare, but that's nice, of course,



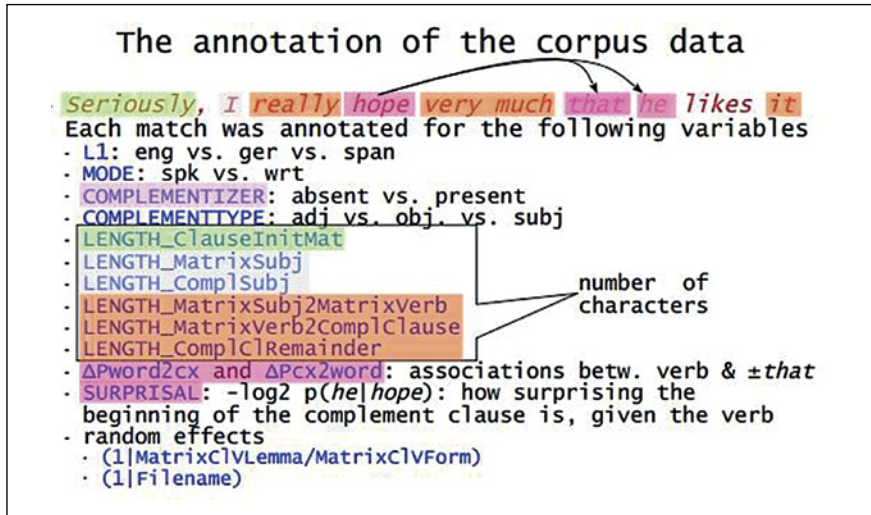


FIGURE 7

because it makes statistical analysis easier. We do have quite a large number of data points, so that's nice and nearly 9500 data points. Just as a general comment, I'm not going to dwell on this much, but as you can see, adjectival complementation in particular by the Spanish learners is extremely rare. So any conclusions about that part would have to be done with a lot of caution, given the literally handful of data points we only have for that. But still, in general, pretty large number of data points. So 9500 lines, what were they annotated for? Well, pretty much all the things you've seen, so that took quite some time.

Again, I'm going to use this sentence as an example to show what was annotated in what way.

Obviously, we annotated the L1, this is where the example was from, so English for the native speakers and then German and Spanish for the learners. Secondly, speaking versus writing, just to see whether that makes a difference, and we will see that it does. Also, of course, because past studies have shown that it has a difference for native speakers, it *makes* a difference from native speakers. Then the dependent variable, or what might be considered at first, at least, the dependent variable, was whether the complementizer was used or not. *Present* means people put *that* in there, *absent* means they didn't. In the example sentence you see here, obviously, it was put in there.

Then the complement type. So is it adjectival or object or subject complementation for every one of those? Then all the lengths of these different kinds of materials, so clause-initial material, the length of the matrix clause subject and the complement clause subject, and the length of any intervening

materials in these three different slots. For ease of processing, those were all operationalized in terms of number of characters, simply because it doesn't make a huge difference which of the length or weight-based operationalization you use so you might just as well go with the simplest one.

Then we computed association scores of the type we talked about, I think, on day three in this week. We computed  $\Delta P$ , word-to-construction and construction-to-word. The arrow here is word-to-construction, right? To what degree does *hope* or to what degree is *hope* attracted to the then following complement clause because different verbs make it differently likely that a complement clause will be following. So we computed that in both directions, again, reminding you of this notion that maybe directional association measures are more precise in some applications.

Then we computed the surprisal of the first word of the complement clause, given the verb. The question basically is, if the hearer hears the *hope* here, how much will they anticipate that the complement clause will begin with *he*? The smaller that probability, the higher the surprisal will be. As I mentioned yesterday, if—what happens a lot of times is that if something upcoming is going to be very surprising, then speakers sometimes have this tendency to insert something that smoothes the processing load so if this thing is really surprising, then people are more likely to put the *that* in because it helps ease the processing of the upcoming complement clause.

Then, as I said, mixed-effects modeling, so we had a random-effects structure, which here was restricted to varying intercepts, namely for the matrix clause, verb lemmas and forms, so *hope*, and for the file name as a proxy for the speaker, to make sure that speakers who never use *that* or who always use *that* get accommodated properly.

Then, when all of these factors were annotated, we did a statistical analysis, but we didn't do what has, until recently maybe, been the standard approach to this. The normal approach for this kind of data set would be to run a regression model, where you try to predict whether people use *that* or not. One of the predictors is the L1, so would be, English versus German versus Spanish. Then, to determine whether all of these factors are related to the L1 of the speaker, any of these would be allowed to interact statistically with the L1. So for instance, to see whether there's an effect of clause-initial material that is different between German learners and Spanish learners, and in turn may be different from the native speakers, one would include the interaction of L1 with this variable in the regression model to see whether that makes a difference. For reasons that are not relevant right now, I developed at some point a different method and that's the one we're using here. It's called Multifactorial Prediction and Deviation Analysis with Regressions [[MuPDAR]]. If you apply

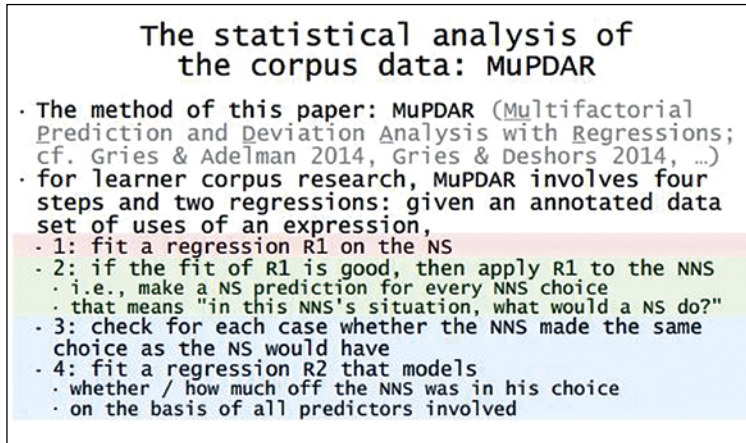


FIGURE 8

it to learner corpus data, then it involves four steps and two regressions. (It can theoretically be run with other classifiers such as random forests or something like that, or classification trees. In this example and the next one, it's always going to be regression.)

So the steps are this: First, you fit a regression model to the native speaker data only, you don't even look at the learner data in the first step. You only run a regression on the native speaker data. Then, you check whether that regression is good, whether it has good predictive power because if it doesn't, you're done with this approach, then you can't do it. But if you have a good regression fit to the data, then you use the regression model run on the native speaker data to predict what native speakers would have done in the situations the learners were in. So what that means is, and this is why it's similar to missing data imputation: Think of it this way, for the learners, we have what they did in a certain situation, because we have what they wrote in the essay and we annotated it. What we don't have for the learners is another column that says whether a native speaker found that choice okay or not. We only have what they did, but we don't know whether that's actually what a native speaker would have done. That's what this step is doing, basically it uses a statistical model trained on native speakers that according to criteria that one might discuss, that according to some criteria is good, and then you use that statistical model to take all of the learner data, and for every one of them say this is what a native speaker would have done here, and this is what a native speaker would have done here and so on. So we're simulating an error annotator as if a native speaker had read all the learner items and had said for everyone, "Yeah, I would have said that", "Yeah, I would have said that", "No, here I wouldn't have" and so on.

So the model answers this question in this learner situation, “what would a native speaker do or have done?” That of course means we can now check for every one of the cases, for every learner choice, whether it’s what a native speaker would have done, which is just a different way of saying, for every learner utterance, we can check whether it’s nativelike or not.

Then we can run a second regression model, namely, where the dependent variable now is a different one. So in the first regression, the dependent variable was is someone using *that* or not. In the second regression model, the dependent variable is whether the learner made a nativelike choice, yes or no? Or if you want to, it’s not anymore politically correct to say this, but it boils down to, “did the learner make a mistake or not?” That becomes the dependent variable. We model that on the basis of all the predictors that we know affect *that*-realization. All this is the complex way of saying this second regression basically figures out which combinations of things are difficult for learners, because when they occur, then the learners make non-nativelike choices. That’s essentially what this is doing. Again: model on the native speakers, apply to the learners, check whether the learners made nativelike choices, and then model what is it that makes learners make non-nativelike choices.

So what are the results? In particular, the overall results or the summary results. The first regression with the native speaker data only contained all the predictors that I mentioned, a bunch of two-way interactions between some of these predictors, and it had a random effect structure that said every speaker is treated slightly differently, and every verb lemma, and every verb form is treated slightly differently. Like I said, you can only use this method if this first

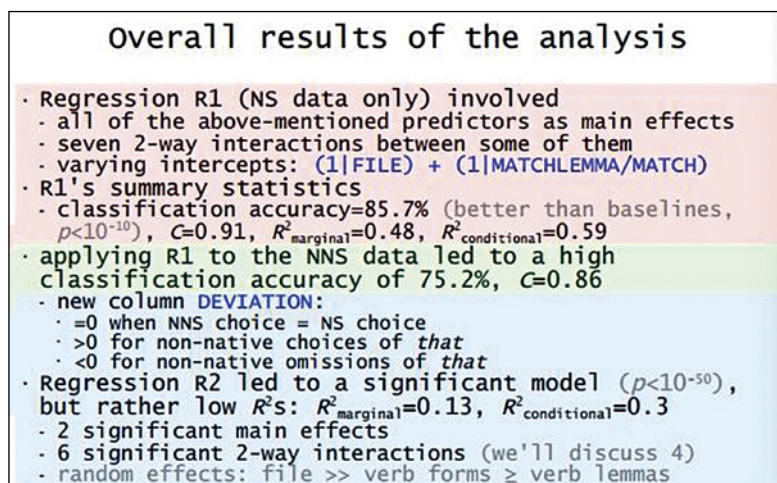


FIGURE 9

regression is good or is successful. If you fail to find out what native speakers are doing, then you can't make native-speakerlike predictions for learner language. So you need to have a good fit here. In this case, that was the case. The classification accuracy (85.7%) that we achieved on the native speaker data was better than chance, quite a bit. The *C*-score that is often used to diagnose regression quality was good as well. You want to see a value of 0.8, we got point nine or more than that even (0.91). The  $R^2$ -values that we have here are not too bad. So we did apply that first regression model to the learner data, and we got a classification accuracy of about seventy-five percent (75.2%), still a good *C*-score, but of course this is doing worse because now we're taking a native speaker trained model and apply it to people who are not native speakers. I mean the very fact that they are learners should lead to a decrease in classification accuracy because learners make mistakes. That's why this number here (75.2%), of course, is lower.

Then we computed a new variable, which is called *Deviation*. Let me explain what that does. If you apply this kind of technique, there are two different ways you can check what learners are doing. One is the one that I mentioned before, namely, you check for every non-native speaker choice whether it's correct or not. That means your dependent variable is binary: learners made the right choice or the wrong choice, simplistically speaking. Now that's relatively crude because there might be cases where even native speakers are like, "Well, you know, I mean I could put *that* in there, I could leave it out", that would be fine. But then if a learner uses *that* there for instance, then that might be wrong, although the native speaker is actually ambivalent about what to do there. Here we did something slightly better. This new column or this new variable *Deviation* was set to something numeric. Namely, it was set to zero when the non-native speaker choice is what a native speaker would have done: If the learner wrote exactly what a native speaker would have written, then there is no deviation, then it's the same so *Deviation* is set to zero.

Then the value was computed in such a way that it would be greater than zero, when the learner put *that* in there and the native speaker wouldn't have. A value of 0.3, for instance, would say, 'the learner put *that* in there, the native speaker would not have and the native speaker was fairly certain of that'. The same thing for less, for values smaller than zero: those would be non-native omissions of *that*. The learner didn't put *that* in there, but the native speaker said, "Yeah, I would have put it in there". You can make mistakes in either way: You can realize the complementizer if it shouldn't go there or the other way around. That variable then became the dependent variable in the second regression. The cool thing about this one now is this is not just right or wrong, but the size of the value tells you how much right or wrong. So a value close to

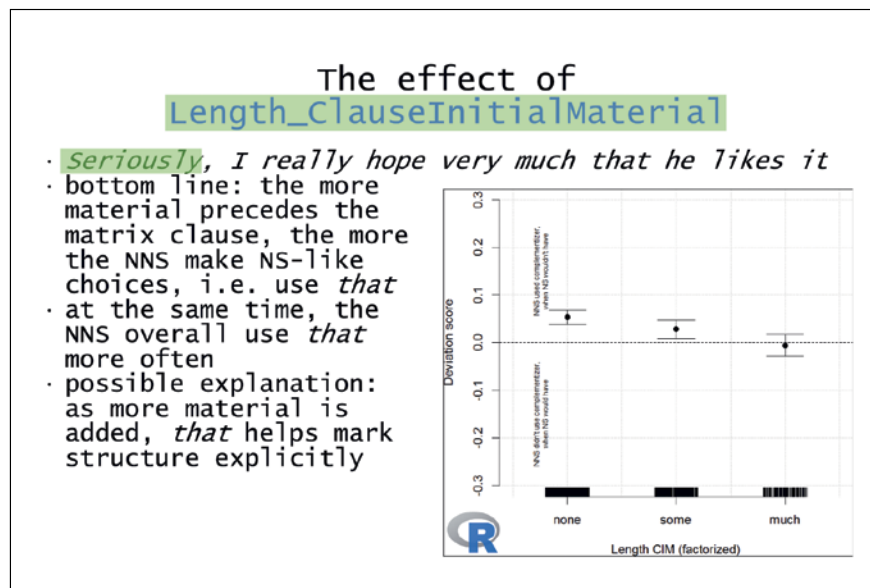


FIGURE 10

zero would mean ‘the learner didn’t do the right thing, but it wasn’t a big mistake’; a value close to 0.5 or close to -0.5 would be like the learner didn’t do the native-like thing. I mean it’s really wrong: The native speakers would definitely not have done that. We can not only see whether there is an error or not, but this deviation column also quantifies the severity of the error.

Then we fit the second regression model to see, can we explain when speakers don’t get it right? We can! The  $R^2$ -values are relatively crappy, but the effects that we’re getting are significant and many of them are pretty well interpretable.

Let me show you some of the effects that we find. One is a main effect, namely the effect of the length of the clause-initial material, the effect of how long this thing is before the main clause. This is the effect. On the x-axis, you have the length of that clause-initial material. For statistical reasons, we had to take that number and force it into one of three groups—*none*, *some* and *much*—here expertly called, there is a principled decision why it’s those three values.

Then let me explain the y-axis because it’s going to be the same everywhere. The y-axis is the deviation score, which means if these predicted values here are in the middle around zero, then that means then the learners get it right. Whereas if the values are higher up than zero or further down than zero, it means the learners did not get it right. Then whether it’s higher up or lower up tells you how they get it wrong. I always have to remind myself, so that’s

why I put this little explanation here. When it's positive, it means the non-native speakers used the complementizer when the native speaker wouldn't have. This is putting it in when you shouldn't [pointing to the upper part of the figure]; this is leaving it out when you shouldn't [pointing to the lower part of the figure]. Right? Again, like here, in the middle is when they get it right. All the other plots on the following slide will be like that.

So what does this show? Actually it shows a very nice effect. Again, *none*, *some* and *much*, so if there's a lot of material here, then the learners make natively like choices, and the natively like choice when there's a lot of material is to put the complementizer in, but they [the learners] get that. But the less material there is, *much* to *some*, *some* to *nothing*, in fact, the more they get it wrong and the way they get it wrong is they put *that* in there when a native speaker wouldn't have. When there's nothing here, if this wasn't there, you know then this could be a learner utterance because it has a *that* there, although it shouldn't have, given that there's nothing here. So no material in front of the clause, and the learners overuse *that*. One reason, of course, might be one that is related to processing, namely *that*, of course, is very useful in how it marks the transition from the main clause to the complement clause. If complexity of the to-be-processed whole sentence is increased, then you know learners, even for themselves, might think or might benefit from flagging the transition from one clause to the other as explicitly as possible, putting the *that* in. Native speakers

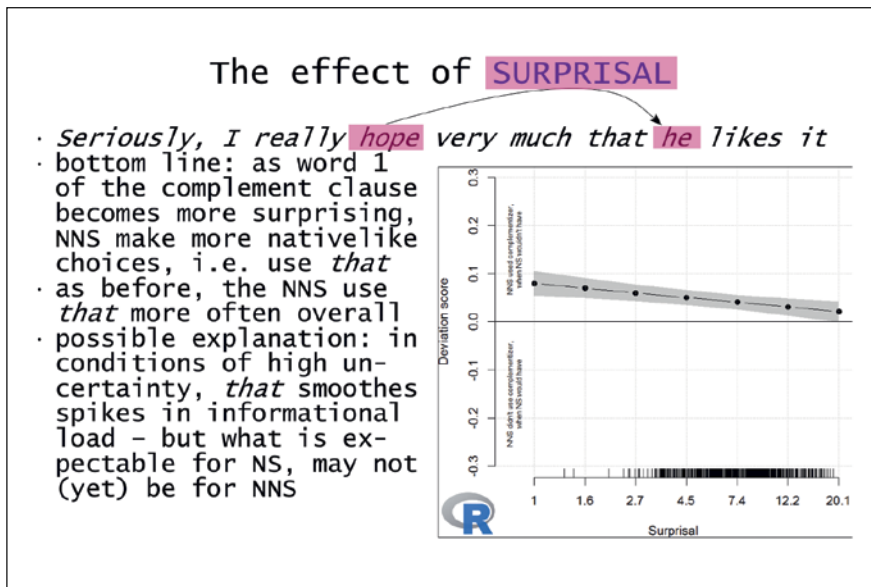


FIGURE 11



will be under a less high processing load because it's their native language, so they don't need that clue of the *that* so much, so they don't put it in. That's why when there's less material, we find overuse of the *that* by the learners.

Secondly and very nicely, especially for this series of talks, there was a main effect of surprisal. The question was to what degree does this verb (*hope*) makes you expect that word (*he*) in this slot? The effect is this: On the x-axis here we have surprisal on a log scale, from very little to quite high. Again, we have a trend that as this thing (*he*) becomes more surprising, the learners become more native-like, the (*Deviation*) values creep closer to zero. Again, they behave like the native speaker, so that means the native speakers, if this (*he*) is more surprising, they put *that* in, the learners do, too. But where the learners differ is when this thing (*he*) is highly predictable, then they still a lot of times put *that* in when they shouldn't, when native speakers would say, "Yeah, okay, this is so predictable. I don't need to smooth this: *that* can go". Part of the explanation is something like, native speakers have had way more input than non-native speakers. So for them, a lot of things that seem unexpected—for the native speakers, a lot of things that seem unexpected to a learner are not unexpected and so that might be responsible in part for this kind of effect.

Here's an interaction. We find an effect such that adjectival, object, and subject complements behave differently, depending on how long this thing (*he*) is, the subject of the complement clause. I'm going to not discuss this one in

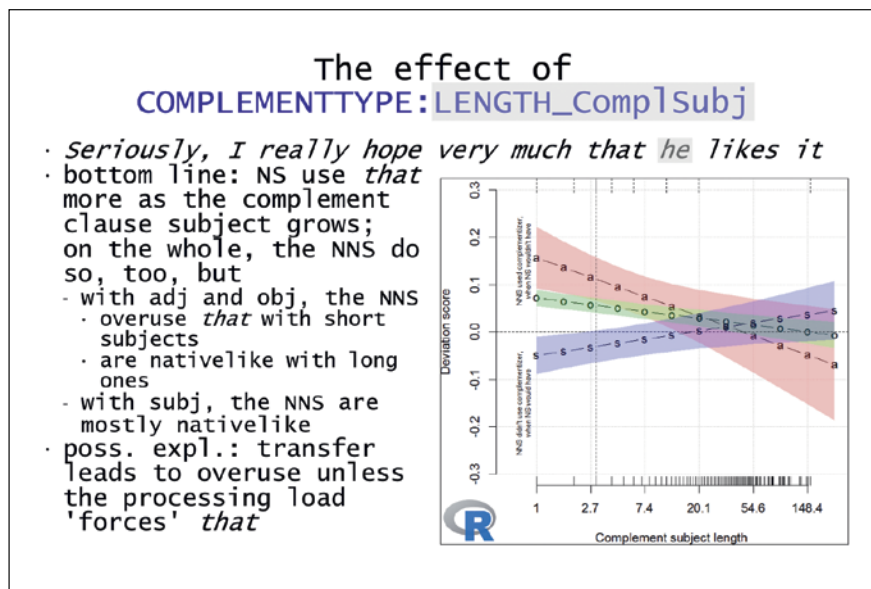


FIGURE 12



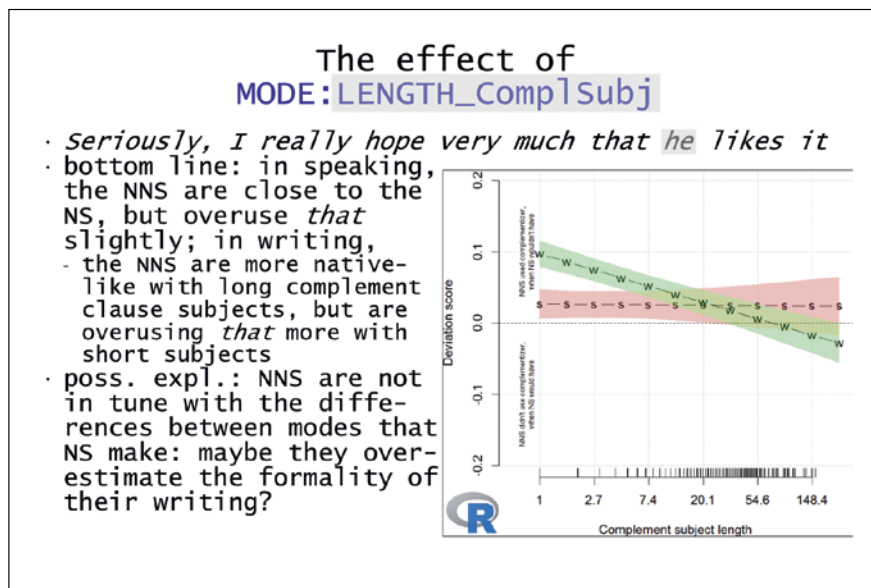


FIGURE 13

great detail. As we can see, with adjectival and object complementation, the non-native speakers overuse *that* when the complement subject is short, right? When the complement subject is short, so on the left side, then the *a* and the *o* curves are highly on top. With short subjects, adjectival and object complementation, they really get it wrong a lot. On the whole with subject complementation, this line is closest to the zero line for the most part. So subject complementation, the learners are best at, compared to the other two. In part, of course, this might be a transfer explanation, in particular for the German learners but I'm not going into that much right now.

We have an effect or an interaction of mode and complement subject length looking like this. This one is interesting: basically in speaking, so the red curve, the learners are close to the native speakers. I mean this line is pretty close to the zero the whole time, but it's also always on the top. They overuse *that* slightly but also the fact that this line is horizontal means that actually the subject complement length has no effect. They overuse it a bit here when the subject is really long, but also here when it's really short. But in writing, it's different. In writing, the learners are more native-like when the complement clause subject is long. When the subject is long, the green line hits the zero, but the learners overuse a lot in writing when the subject is short.

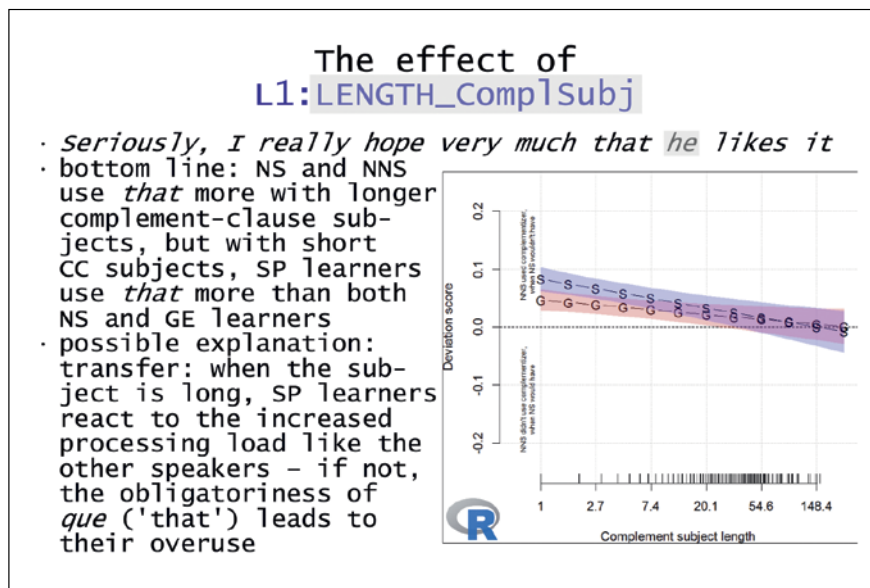


FIGURE 14

Let me briefly talk at least about one effect with L1, and probably in the interest of time, I might skip over the other. Let me know if you want this discussed later. There was an interaction, a difference between the German and the Spanish learners when it came to the length of the complement subject. So how long this thing (*he*) is? Of course I have great pride and pleasure in reporting that the German learners did better than the Spanish learners. There's, again, the general pattern of overuse: The errors are on this side of the plot [pointing to the upper part of the figure], so the learners put *that* in more often than they should, but especially with the short complement clause subjects, the Spanish learners overuse more than the German learners overuse, and then they're relatively comparable in their overlap here at the end. Of course, that might be in part a factor of processing, that again, people put the *that* in here when the subject is long, so that leads to an issue. Then in the short part here, where processing load is probably not the explanation, it could be the obligatoriness of *que* that leads to the Spanish language users overusing the *that* compared to the German ones.

Again, I'm going to skip this one in the interest of time.

So what do we find? On the whole, the learners are doing pretty well. Applying the native speaker trained model, we did get seventy-five percent

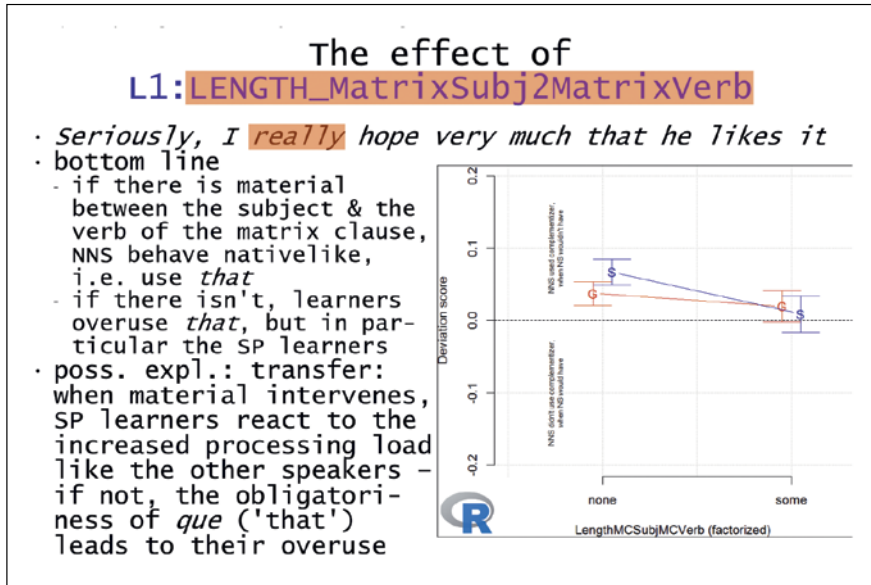


FIGURE 15

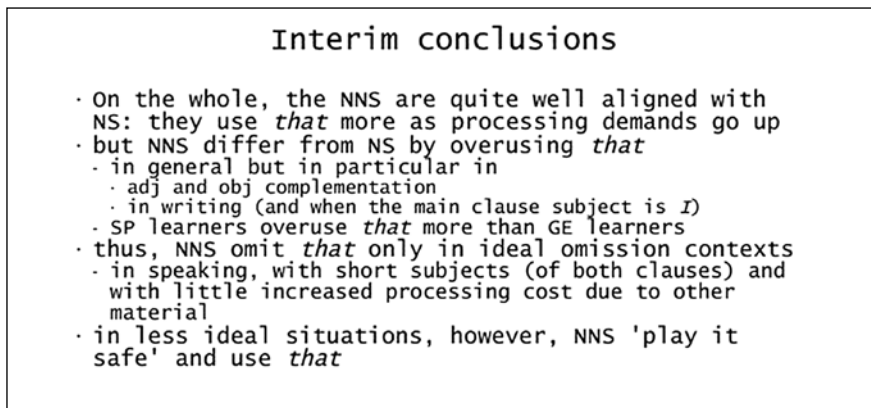


FIGURE 16

right. So these learners, they are a sort of intermediate to advanced learners in this corpus, on the whole, they're doing pretty well. The general thing they do is exactly what native speakers do, namely when processing demands go up, because things get longer, things get more complex, more material is integrated and everything, then they put in *that* more, just like the native speakers. But they differ with a general overuse of *that*, and they do that in general, but

particularly in adjectival and object complementation, they do that also particularly in writing, and they do that when the main clause subject is *I*, where especially in speaking native speakers really don't use *that* very much at all. These are the kinds of settings that lead to particularly non-nativelike behavior. The Spanish learners overuse *that* more than the German ones—that's of course great. So essentially what happens is that the non-native speakers omit the *that* in ideal omission contexts, when everything says, 'this is all easy, this is all short', then the learners omit the *that* as well. That would be, in speaking, it would be with short subjects of both clauses, and when there's no other intervening material between the subject and the verb, the verb and the *that*, and the *that* and the subject: whenever there's nothing going on in there, then the learners are happy to omit the *that*, but otherwise they put it in, they play it safe and use it. Of course, it's not usually a mistake. I mean just because you put *that* in where it's optional doesn't make it some mistake. I mean they are avoiding grammar errors, and many of these settings were, of course, they're graded, but still, it's kind of non-native.

So now the nice thing then is that the statistical approach could do a lot of nice things. Remember that at the beginning of the week I told you overall aggregated frequencies of stuff—so how much do learners in general use something? how much do learners in general not use something?—as soon as you aggregate over speakers, over files or something like that, you basically lose a ton of information.

### Concluding remarks: summarizing the results

- On the whole, the NNS are quite well aligned with NS: they use *that* more as processing demands go up
- but NNS differ from NS by overusing *that*
  - in general but in particular in
    - adj and obj complementation
    - in writing (and when the main clause subject is *I*)
  - SP learners overuse *that* more than GE learners
- thus, NNS omit *that* only in ideal omission contexts
  - in speaking, with short subjects (of both clauses) and with little increased processing cost due to other material
- in less ideal situations, however, NNS 'play it safe' and use *that*
- the MUPDAR approach has offered us a picture that not typical (aggregating) over-/underuse kind of study can readily provide (in terms of comprehensiveness and statistical control of competing factors)  
(see Lester 2018, *IJLCR*, for a similar study)

FIGURE 17

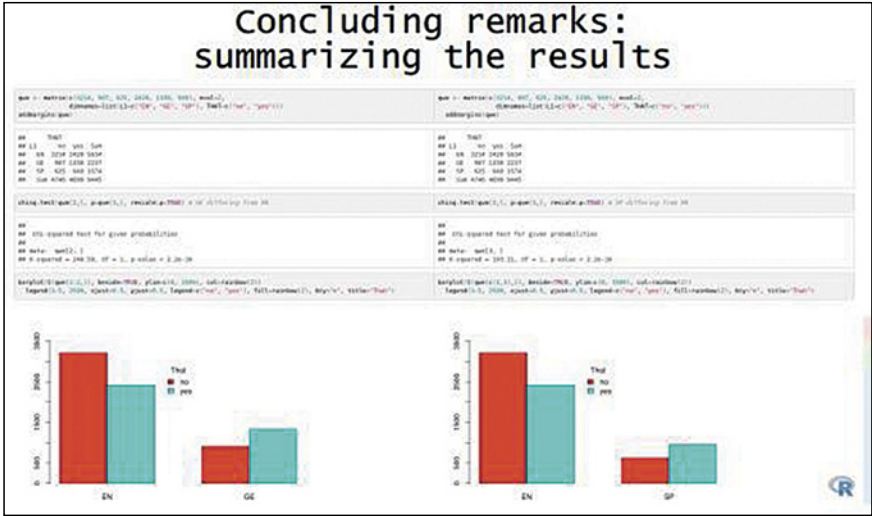


FIGURE 18

I just want to show you here what the traditional learner corpus research analysis would have shown. In a traditional approach, and there's a ton of work out there that does it like this, people would have done something like this: It's probably too small to read, but just to give you an idea, as they would have created a table that has three rows for English, German, and Spanish, and that has two columns for complementizer, no and yes. Then they would have compared the frequencies with which English speakers use or omit *that*, so that's the left part here, there we go, and they would have compared it to the German learners, which is this part here. Then, there is a lot of work that then just runs a chi-squared test on that, and of course finds that, the German learners overuse it because there's more of the blue stuff where *that* is used compared to here where *that* is less used. Then, a lot of traditional learner corpus research would do the same for the Spanish learners, so here on the right side, the same distribution for the English data as here and then, for the Spanish learners, we have this. Again, we have an overuse of *that* compared to here. That would be it. I mean, there's a whole bunch of papers that basically do that. I dare say you know that the analysis that we did is way more comprehensive because we could tease apart what all these individual factors are doing, how they interact with language, I mean L1, and how they interact with mode and all these other kinds of things. So aggregating frequencies, reporting something for a corpus as a whole, like the whole Spanish learner corpus, it doesn't make a lot of sense, especially not if you want to be cognitively realistic in any meaningful way.

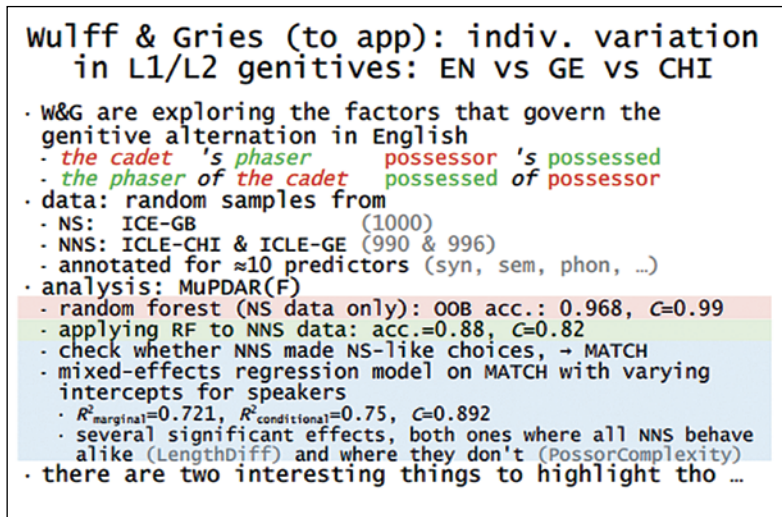


FIGURE 19

The second case study, this one will be relatively brief. It's a very similar approach. The main focus here is going to be on the analysis of individual-variation data. But again, otherwise it's relatively similar. We looked at genitives in L1 and L2, so the *of*- versus the *s*-genitive alternation, so *the President's speech* versus *the speech of the President* or something like that.

We have native speaker data and we had data from German and Chinese learners of English. The question is, what are the factors that make people choose either one of these two [pointing to the two examples]? Note here in particular, you know that this alternation is tricky, because you know it doesn't just involve the change from *s* to *of*, but also the flipping of *possessor* and *possessed* in the order. The *s*-genitive has the *possessor* first, right? This person possesses this concrete object. The *of*-genitive has the possessed thing first, the thing that is owned by, in this case, the cadet.

We had random samples from a variety of different corpora, specifically again, the British Component of the International Corpus of English, and then ICLE Chinese and ICLE German, roughly pretty much exactly 1000 items for each. Here obviously, we couldn't look for all of them because the *of*- and the *s*-genitive are just so frequent that we did not want to spend the time on annotating that. Then, we had approximately ten different predictors from a variety of different levels of linguistic analysis, so syntactic ones, semantic ones having to do with the semantic relationship that the genitive encodes, but also phonological variables actually, which is interesting because—maybe as a short

side remark: the data we looked at here are written data the ICLE corpus, again International Corpus of Learner English, that's written data for the Chinese and for the German learners, but still actually in the analysis, we found results that I'm not going to discuss here, but that showed that there's a phonological effect that you find even in writing, right? So a lot of times people are like, "Well, the corpus data are written data. Is that also true in speaking?" Or "I bet it would be totally different in speaking"—actually, if you do the analysis, you often find traces of phonological or articulatory effects *even in writing*. We found that here, too. So don't let people shoot your work on written data down just because, you know, it doesn't contain anything about spoken data, a lot of times you might have effects in there as well.

We did the same kind of two-step analysis. In this case, we used a random forest analysis on the native speaker data. We got an absolutely crazy classification accuracy (96.8%), like nearly 97 percent right. That's even a prediction accuracy, not just classification, but prediction. So really good. We did apply that to the learners. We still got an extremely good classification accuracy. Then we looked at, okay, did the learners make native-like choices, yes or no? We did a mixed-effects regression model on that. We got a bunch of significant effects out of there that we could explain relatively well, but the main point *here*, again, is that of individual variation.

That is interesting because in most traditional over-/underuse studies, the ones that conflate across the whole corpus, you of course don't have that: remember these two red and green or whatever bar plots that I just showed you,

### Wulff & Gries: 2 interesting aspects

- 1: another bit against traditional over-/underuse
  - we found a significant interaction effect such that
    - NNS make very nativelike choices when POSSOR is animate
    - NNS make very non-nativelike choices when POSSOR is inanimate & POSSED is plural
  - but the traditional  $\chi^2$ /LLR test of these frequencies is ns ( $p > 0.95!$ ) because this aggregates and does not control for anything else in the data ...
- 2: there is substantial individual variation
  - re % of non-nativelike choices
  - re degree of non-nativelike choices
  - this variability
    - is not captured in most approaches that aggregate
    - is not easily isolatable when we do not control for other predictors
    - can be correlated with speaker-specific information both *a priori* or *a posteriori* / exploratorily
    - even allows to explore individual non-nativelike decisions and the factors that motivated them

FIGURE 20



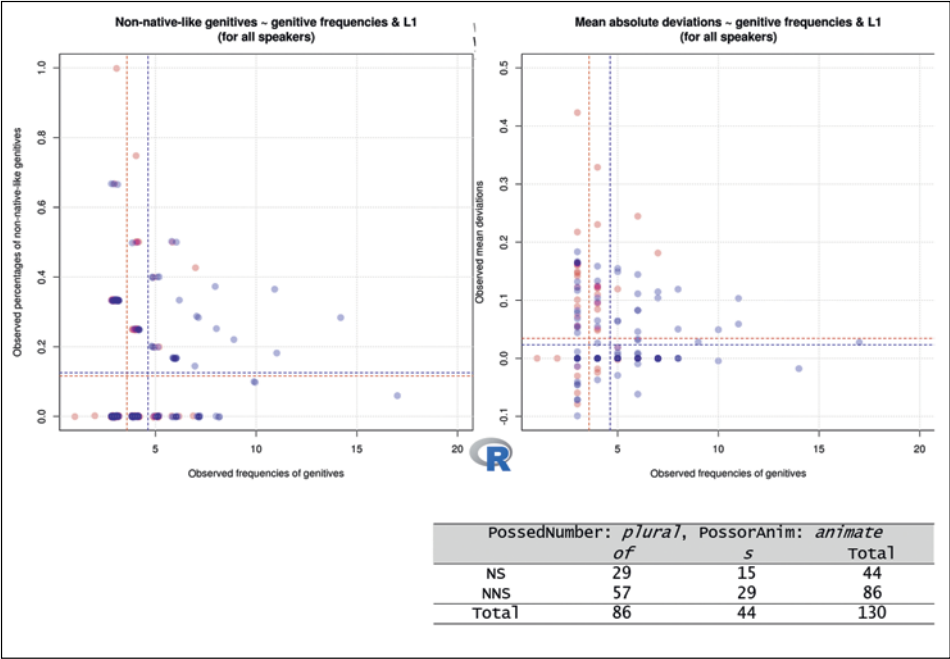


FIGURE 21

I mean, they just amalgamate everything from every speaker without looking at what happens on a speaker-by-speaker basis. You don't see any individual variation effects, but we actually did find effects that highlight, on the one hand, the fact that many of the predictors we were looking at are interacting with each other, but secondly, and this is the part I want to focus on, that there is a huge degree of individual variation: Learners have their pet constructions that they fall back on by default. Of course, learners often don't maybe even know that an alternative construction exists. This is a case where individual variability is really important to look at. What we found is that speakers differ a lot, both in terms of how often they get it right, but also in terms of how wrong their choices are. Some people are off a little bit, they use an *s*-genitive when actually both would be acceptable, but a native speaker would have slightly tended towards an *of* genitive, and some really get it wrong.

Here's one visualization of something like this, not to look into this in too much detail: every one of these blue and red points is a speaker. I think blue are the German ones and the Chinese are the red ones. Then on the *y*-axis here, we have the percentage of how often do people make mistakes. Here's someone who gets it wrong all the time, and then here is how severe the errors are. Then, the red and the blue lines are the medians. So for instance here, we



## wulff & Gries: 2 interesting aspects

- 1: another bit against traditional over-/underuse
  - we found a significant interaction effect such that
    - NNS make very nativelike choices when POSSOR is animate
    - NNS make very non-nativelike choices when POSSOR is inanimate & POSSED is plural
  - but the traditional  $\chi^2$ /LLR test of these frequencies is ns ( $p > 0.95!$ ) because this aggregates and does not control for anything else in the data ...
- 2: there is substantial individual variation
  - re % of non-nativelike choices
  - re degree of non-nativelike choices
  - this variability
    - is not captured in most approaches that aggregate
    - is not easily isolatable when we do not control for other predictors
    - can be correlated with speaker-specific information both *a priori* or *a posteriori* / exploratorily
    - even allows to explore individual non-nativelike decisions and the factors that motivated them

FIGURE 22

can see that the German learners on the whole used genitives a little bit more often than the Chinese ones, but the Chinese ones made slightly more severe mistakes in the genitive choices. But the main point is actually not that—the main point is the relatively big amount of scatter that you see that indicates a lot of speakers are really different from each other, both in terms of how many genitives they use—look, there's some people here who use them a lot, many people use them rarely—and then also in terms of how wrong they go: some people here are wrong a lot of the time and pretty severely, many people are on the zero line, so they get it right all the time. You cannot see that if you aggregate frequencies without any regard as to where the data come from.

In this case, we didn't do that in great detail, but one thing you *can* do is you can look at these kinds of plots, you can look at the individual speakers that are represented by them, and then see whether anything else that you know about the speaker correlates with their degrees of mistakes or their numbers of mistakes. We didn't do this, but for instance, theoretically, if we had the data, it would be interesting to see. Are there speakers that spend a year or half a year in an English speaking context? Are those the ones that get it right more often? Are those the ones that make less severe errors or something like that? As soon as you look at the level of individual speakers, of course you can use many more explanatory mechanisms, in order to try to make sense of the data. So again, that's kind of the plug for why we think it is so important to include this kind of information in there.

## Interim conclusions

- Much existing work underutilizes data (incl. my own)
- many over-/underuse studies are
  - useless when they aggregate & involve  $\chi^2$ /LLR tests
    - control for nothing, disregard dispersion, explain next to 0 variability statistically & conceptually, ...
  - better when they are reconceptualized as glm(er)?s
    - two levels of resolution: by file/speaker or by phenomenon
    - captures some speaker variability, integrates methods nicely
  - but even those are not helping much: they still control for nothing (else) & explain little (interesting)
- we need (something like) MuPDAR(F)
  - detailed case-by-variable annotation for every instance of x and ever instance of it before it (priming):
    - sampling level of speaker/conv (rep meas, order/learnng ...)
    - speaker-specific predictors (cogn, psych, proficiency, ...)
    - linguistic predictors (syn, sem, inf-structural, phon, ...)
    - textual & task predictors (genre, lex diversity, topic, ...)
  - detailed! comparisons of NNS to NS production
- you cannot study overuse w/out asking "why/where?", which means you must control those statistically!

FIGURE 23

A lot of the time, again, as a bottom line here, the kind of statistical structure that we do find in these kinds of heavily-annotated data is often underutilized. If you aggregate, you basically control for nothing, you disregard dispersion across different speakers, and actually, if you adopt a regression-like context, you actually explain pretty much zero percent of the statistical variability in the data. It's really useful for these kinds of contexts to study this with a logistic regression modeling context, especially as a mixed-effects model like some of you that I've talked to are already doing, because you do want to be able to distinguish between different speakers, between different lexical items, and all these kinds of things that might have a huge impact.

If we do run an analysis like the one that I've been talking about here, however, we also find that a lot of corpora that we're using are lacking in many respects. This is not to criticize the corpus compilers, obviously, those are huge projects that involve a lot of hours and stuff like that. But what is routinely not studied in great detail is, first, something like priming, and we talked about this in the recency secession but then also, what we ideally would have, in particular for a second language acquisition or learner corpus research studies, is more speaker-specific information. Some corpora offer information on proficiency scores or something like that but we know next to nothing for most speakers about things having to do with personality, aptitude, motivation, all these kinds of things that are ultimately strongly correlated with learner proficiency and success. If we want to make this kind of corpus work more relevant

## The final final case study

- For a theory that claims that
  - linguistic knowledge is knowledge of constructions
  - linguistic structure/representation emerges from use
- maybe obvious questions are
  - whether NNS build up their constructicons as NS do
  - whether non-native speakers' constructicons are different from those of native speakers
  - are differences quantitative tendencies or qualitative?
- this paper looks at the **dative alternation** (with *to*)
  - *Picard gave* [<sub>NP REC</sub> *the Borg*] [<sub>NP PAT</sub> *his phaser*]
  - *Picard gave* [<sub>NP PAT</sub> *his phaser*] *to* [<sub>NP REC</sub> *the Borg*]
- specifically,
  - do German learners of English exhibit **structural priming** effects comparable to English NS?
  - what affects priming effects of German NNS?
  - are priming effects of NNS correlated with the **verbs' distributional preferences** in NS corpus data?

FIGURE 24

to an SLA audience or to an FLA audience, then I think corpora in the future will probably need to do a little bit more in terms of adding information about individual speakers to the mix. Alright, this one I'm going to skip.

The final final case study, at long last ... The usage-based model, like I discussed yesterday in detail, assumes that linguistic knowledge is knowledge of constructions (it's constructions all the way around from the bottom to the very top) and that linguistic structure and representation emerges from use. Then obviously questions would be, how and whether non-native speakers build up constructicons in the same way that native speakers do. To what degree does a non-native learning context or acquisition context lead to the same kind of constructicon as that what native speakers build up during the first years of their lives? Then what kind of differences can we see there? Are those quantitative differences? Are those completely qualitative differences? Obviously, for a usage-based approach, these things would be relevant. So what I want to do here in this paper or case study is look at the dative alternation, the alternation between the ditransitive construction here in the first one and the prepositional dative with *to* in the second one and specifically look at whether learners of English exhibit priming in a way that is comparable to that of native speakers. We've talked about priming multiple times during this week because it is such an important and such a good diagnostic of mental representation. So, obviously, if you assume that learners build up a constructicon, then that kind implies that you would expect to see structural priming effects there as well: If they represent a construction mentally, then it should be possible to prime it and make them use it again, compared to a contrast or control condition.

### Previous work on this

- Gries & Wulff (2005) replicated a series of priming experiments from Pickering & Branigan (1998) and indeed found a priming effect
  - odds ratio for priming: 2.57 (95% CI: (1.85, 3.58))
  - $\chi^2=34.55$ ,  $df=1$ ,  $p<0.0001$
- also, they found an overall correlation between
  - the verbs' constructional preferences in the NNS' sentence completions &
  - the verbs' constructional preferences in NS corpus data
- however, according to today's standards, their statistical analysis was insufficient
  - they conflated prime & target frequencies from different experimental conditions for one overall test (Bock 1986)
  - they did not account for speaker-specific variability in the data
  - they did not control for within-experiment learning/habituation effects
- so, let's do better than that ...

FIGURE 25

Secondly, we want to look at, in this particular case, what is it that affects priming effects in turn? So priming effects are not going to be uniformly strong in every different condition. What are the kinds of things that might happen or that might make priming stronger or less strong in a non-native speaker construction? In particular, what we're interested in is, of course, [...] do the learners pick up the distributional tendencies or distributional preferences of verbs in native speaker corpus data? If you're an advanced learner, you've probably had a lot of input, in this case, from English so does that lead to, like an exemplar-based approach would predict, a similar kind of verb-specificity of verb-constructional preference, as that of a native speaker? Especially maybe to one that is different from what the learners do in their native language. That's actually what we want to look at here.

This goes back to a study from 2005, where we replicated a priming study and we did a relatively simplistic analysis that showed that we had a priming effect in German learners of English dative alternation. There was an overall correlation between the verb use of the learners in their sentence completions and what these same verbs do in native-speaker corpora in English. The relatively advanced learners of English here sort of did in fact exhibit the same kind of preferences, but the statistical analysis we did at the time was really not that great. So, for instance, what we did here is we glossed over distinctions between prime and target frequencies from different experimental conditions. Back then, we did not look at speaker-specific variability in the data, and at the time we didn't control for within-experiment learning, so that's what we want to do now to get a better analysis running on this data set.

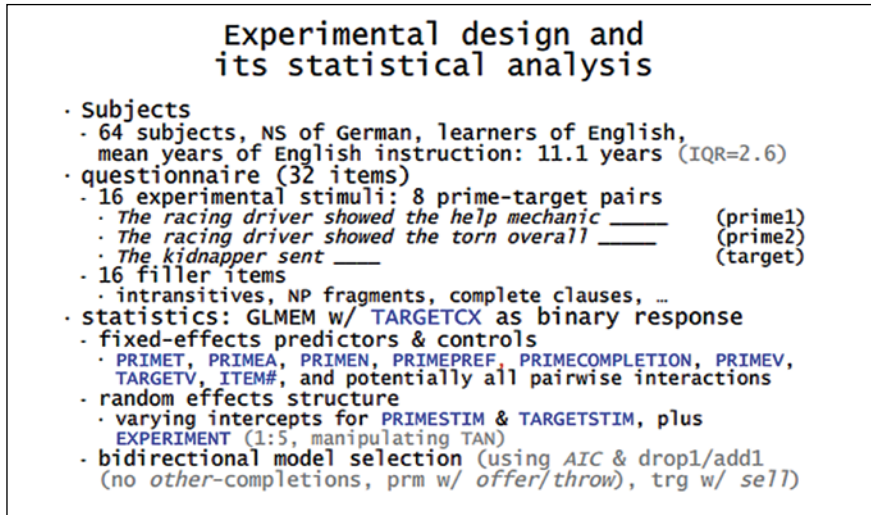


FIGURE 26

What's the design? We had 64 speakers, all native speakers of German, all learners of English with a relatively high number of mean years of English instruction, eleven years. They were given a questionnaire with 32 items, sixteen of these were experimental stimuli, the others were filler items, and the experimental stimuli were prime-target pairs. The way this works is as follows: A subject would get this sentence, *The racing driver showed the helpful mechanic* \_\_\_\_\_ and then they were asked to complete the sentence. So prime<sub>1</sub> is a really strong invitation to complete this in a ditransitive way. *The racing driver showed the mechanic the broken car* or something like that. The second prime is a huge invitation to complete this with a prepositional dative. *The racing driver showed the torn overall to the mechanic*, right? What we did is we gave people one of those two priming sentences, which would bias the learners to complete the sentence in either this way or that way, not that they always did it, but they were strongly nudged in that direction, and then they got a target sentence which ended after the verb, so here you can do both a prepositional dative and a ditransitive equally well. So the question was, if they see this one and complete it with the ditransitive, do they then do that again here [pointing to the target sentence]? If they see this one [pointing to prime<sub>2</sub>] and complete it with a prepositional dative, do they do that again when they see this one [pointing to the target sentence]? Basically, subjects were functioning as their own primes, and then completed the sentence with the target sequence. So the better statistical analysis now uses a mixed-effects model. The dependent variable



## Corpus-linguistic design and statistical analysis

- Corpus data
  - British Component of the International Corpus of English
    - 60% spoken, 40% written
  - 1035 ditransitives, 1919 prepositional datives
  - for each verb attested in the dative alternation, I computed an association measure (AM) quantifying how much the verb 'likes' the prep. dative:  $\Delta P_{\text{PrepDat|V}}$
  - differences to nearly all traditional work?
    - most AMs are bidirectional
    - most AMs combine association & co-occurrence frequency
  - I then correlated the  $\Delta P_{\text{PrepDat|V}}$ -values with the predicted probs of prep. datives from the model in the experiment
    - measure: Spearman's  $\rho$

FIGURE 27

is the target construction, so the construction that is used here after *sent*, while we only use the cases that were prepositional datives or ditransitives.

We had a bunch of predictors and control conditions. In particular, what is relevant, the prime construction, the construction that they actually put in there, the construction that this was priming you in the direction of, the prime completion, so what did they actually put in there? Which verb is being used here? What is the verb in the target fragment? Then for within-experiment learning: the item number? How far along the experiment have they been when they completed this one [pointing to the target sentence]? Was it still at the beginning of the experiment, was it in the middle, was it in the end, or what not? Then we had random effects, namely varying intercepts for every stimulus, but also varying intercepts for every experiment, because we put all the experimental results from five different experiments together that manipulated tense, aspect and numbering in the stimulus sentences.

What did we find? In the corpus data that we wanted to correlate the experimental data with, we did a  $\Delta P$  type of analysis. We looked at every verb that is used in a ditransitive or prepositional dative and we computed a  $\Delta P$  value that says how much a verb likes to occur in the prepositional dative. This is interesting because like I've said before now, during this week, most association measure work is bidirectional—we are separating this now. Most association-based work combines association and co-occurrence frequency—we're using  $\Delta P$ , so we don't conflate. We're trying to keep everything as clean as possible, for when we relate the experimental data to the corpus data. Then, after we were done with the statistical analysis of the experiment, then the predicted probabilities for all the verbs to use the prepositional dative were correlated

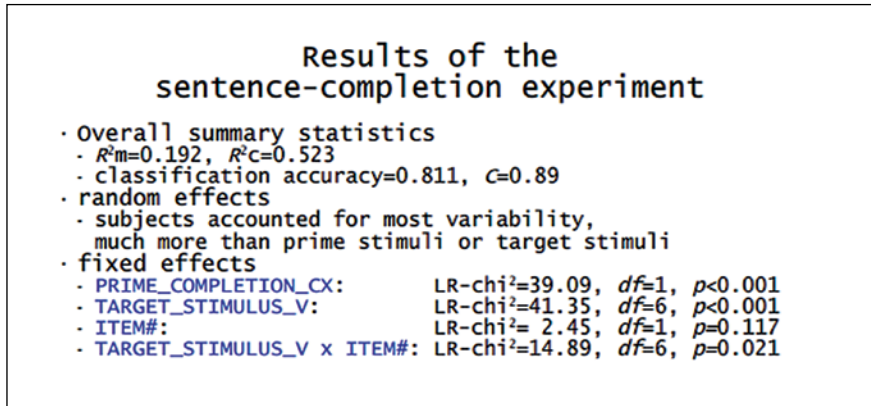


FIGURE 28

with the native-speaker association measures. So the question is in this better reanalysis of the data, do we still find that the learners are behaving relatively nativelike?

So in the sentence-completion experiment, the structure in the data was relatively well identifiable: We did get a good classification accuracy, so there is something going on. There was very relatively little variability across stimuli, but there was a lot of variability between the subjects. The speakers differed from each other a lot, the reactions towards the stimuli did not. We had a bunch of significant effects here, and the strongest main effect is this one, the construction in the prime completion. The strongest main effect was that the learner did seem to use the same construction they used last time, if they could: The completion of the prime had a strong impact on the completion of the target.

There was another effect, namely that of target stimulus verb: Which construction does the verb in the stimulus like most? But, we can't interpret that because it is interacting with the item number. The effect of the target verb changes over time in the experiment: At the beginning of the experiment, learners react differently to something like *The kidnapper sent* than they do at the end when they've already seen a bunch of stimuli. That's the within-experiment learning effect that I was alluding to earlier.

This is the main effect of prime completion. It's pretty strong and pretty straightforward. Here you have the construction of the prime that was used in the prime: Subjects completed something with a ditransitive or with the prepositional dative. Then on the y-axis here, we have the predicted probability of a prepositional dative. You can see if subjects completed something with a

## The main effect of PRIME\_COMPLETION\_CX

- The effect of this predictor **PRIME\_COMPL\_CX** is straightforward
  - when the subjects completed a prime with a prep. dative, they are much more likely to also complete the target that way, and vice versa ( $OR=4$ , 95%  $CI=(2.4, 6.18, nsim=50)$ )

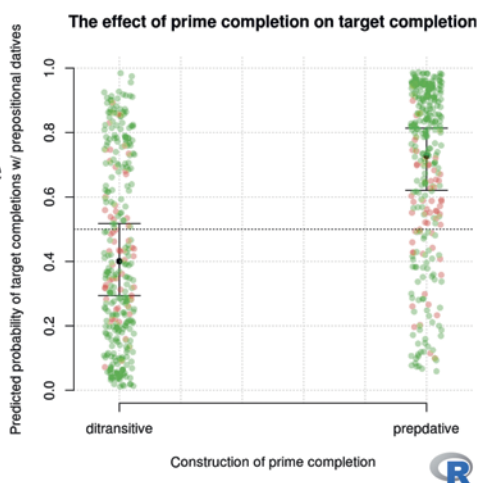


FIGURE 29

prepositional dative, they were more likely to do it again. If subjects completed something with a ditransitive, they were more likely to do it again. We did find the expected self-priming effect with a relatively strong odds ratio, so quite a bit of an effect.

I want to skip this one because we have the interaction effect. This is the main effect, but it was qualified by the learning effect and that's what we see here.

Over the course of the experiment, the subjects' completion preferences change. We had eight different verbs. As the experiment went on—here you have basically the time axis of the experiment without the fillers—then, for instance, people started out with *give* being solidly ditransitive only, and then over time they relaxed that, and were more likely to use it as a prepositional dative. So one thing that's kind of tempting here—I'm not sure I want to speculate on it too much—but obviously, it looks like things start from more extreme values and then kind of converge in the middle. That is interesting or funny in the sense that, of course, this was a sort of nicely-designed experiment where all the stimulus verbs were shown equally frequently in each construction. So at the beginning, they started out with distributional patterns that are pretty much like what happens in native-speaker corpora, but one and a half dozen



## The main effect of TARGET\_STIMULUS\_V

- The effect of this predictor is a bit more complex
  - *give* has the highest dis-preference for prep. dat.
  - only verb coming close to that: *show*
  - verbs 'liking prep. dat.': *post*, *send*, *loan*
  - these results support iconicity analyses (e.g., Thompson & Koide 1987)
- however, note that this predictor was actually part of an interaction with ITEM#

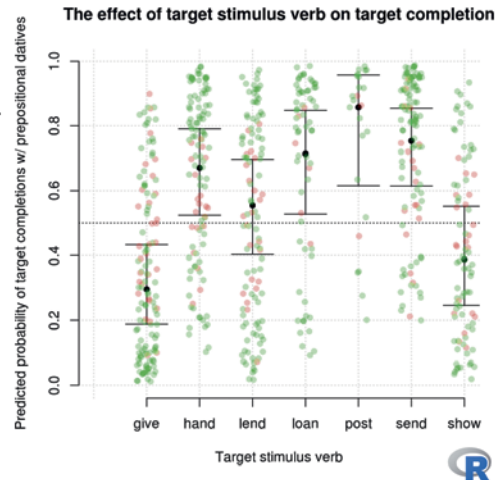


FIGURE 30

## The interaction of TARGET\_STIMULUS\_V x ITEM#

- Over the course of the experiment, the subjects' completion preferences change
- this shows how useful such controls are even/esp. in carefully planned experimentation

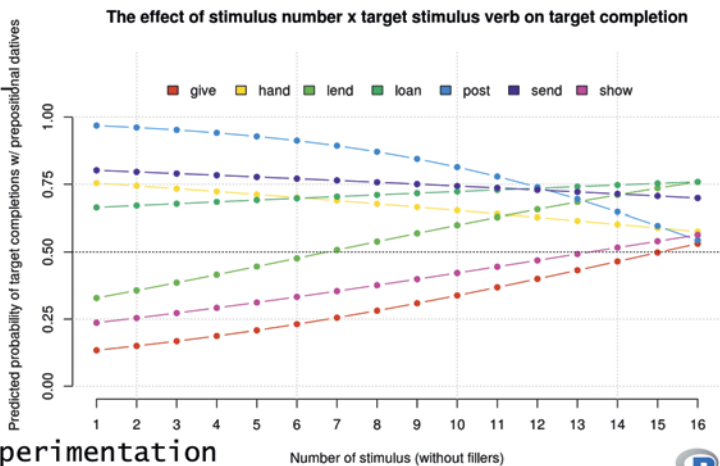


FIGURE 31

## How do these NNS experimental results relate to the NS corpus data?

- For each verb,
  - the experimental NNS results provided us with
    - observed percentages of prep. datives in the completions
    - predicted probabilities of prep. datives from the model
  - the corpus NS data allowed us to compute  $\Delta P_{\text{PrepDat|V}}$
- the observed pairwise correlations are really high
  - Spearman's *rho* of obs. perc. of prep. datives with  $\Delta P$  is 0.9 ( $p_{1\text{-tailed}}=0.007$ )
  - Spearman's *rho* of pred. prob. of prep. datives with  $\Delta P$  is 0.83 ( $p_{1\text{-tailed}}=0.029$ )
- the German learners' overall completion preferences are strongly correlated with the English verbs' subcategorization preferences

FIGURE 32

items later, when they've seen a certain distribution, everything becomes much more evenly distributed, which of course is exactly the distribution that you have in a balanced experiment—everything is equally frequent in each condition. So it is tempting to consider this trend towards the mean here in that way, although I'm not a 100% sure that that is in fact the right explanation.

Now how do these experimental effects relate to the corpus data? Again, maybe in the interest of time, let me just show you the blue summary here: With different correlation coefficients and the  $\Delta P$  values, there's a very, very strong correlation between what the learners are doing and how those very same verbs behave in native speaker data. As you can see here, the correlation of the observed percentages of prepositional datives with the  $\Delta P$  value from the corpora is point nine (0.9), obviously extremely high. In this case, the German learners are pretty advanced, they do exhibit the same distributional patterns as the native speakers do.

Let's wrap up. First, with regard to this study: do the learners' constructional choices exhibit the same kind of preferences? Yes! There is significant production-to-production priming. The strength of the priming is comparable to that of one of the earliest groundbreaking studies on syntactic priming and in a follow up study from a few years later, we had similar kinds of effects for *to/-ing* complementation so there really does seem to be an effect like that in advanced learners.

What is it that affects the non-native speakers' priming? Well, actually a whole bunch of things: Prime-related facts, what did they do last time? But also

### Interim conclusions 1: the research questions

- Do NNS constructional choices exhibit the same kind of priming effects as NS?
  - there is significant production-to-production priming
  - the strength of the priming is  $\approx$  that of Bock (1986)
  - Gries & Wulff (2009) also report priming effects for to/-ing complementation patterns/constructions
- what affects the NNS' priming?
  - both prime-related (PRIME\_COMPLETION\_CX) and target-related (TARGET\_STIMULUS\_V) predictors affect priming - comparing the effect sizes is not straightforward, though, because of the interaction w/ ITEM# (also, see the correlation with the NS corpus data)
- do the NNS exhibit NS-like verb-subcat. effects?
  - yes, NNS' completions are correlated with NS verb-construction associations (esp. more w/ the directional measure  $\Delta P$  than w/  $p_{VE}$ )

FIGURE 33

### Interim conclusions 2: on methodological triangulation

- The results hold valuable lessons
  - experimental designs need better statistical analysis than is often employed
    - multifactorial, random-effects structures, statistical control for learning/habituation,  $R^2$ 's are not always used
    - although we know how many factors can conspire and thus need to be controlled
      - for instance, priming is also affected by prime-target similarity, surprisal, and others (not \* here)
    - although we know how quickly subjects learn in an experiment
  - corpus studies need better statistical analysis than is often employed
    - if an experimental design employs a V-Cx direction, maybe one's AM should, too
    - it is necessary to also always at least consider keeping frequency & association/contingency separate
  - methodological triangulation can be useful (duh)
    - esp. since the control of experimental data poses problems
    - esp. since the noise of observational data poses problems
  - we need both!

FIGURE 34

target-related facts, which verb am I supposed to complete a sentence with? All seem to have an effect. But there also is the learning effect over the course of the experiment. Then, like I said, yes, the learners do exhibit the same kind of preferences as the native speakers do.

In terms of methodological triangulation, this is interesting because as I've mentioned before during this week: experimental work is all great—I've done a lot of experimental work myself—but much of it needs better statistical

### The final final! conclusion

- Corpus data have a lot to offer ...
- a lot more than most ppl give them credit for ...
- frequencies, yes, but also dispersion, association/contingency, surprisal, variability, prototypicality, ..., and often ecologically valid
- but with
  - this richness of information
  - the noisiness of observational data
- comes responsibility
  - knowledge of what corpora can and cannot do
  - statistical expertise
- comes the power to develop, test, refine, and bury theories (esp theories claiming that language/grammars are the product of experience!)
- while some may be too scared of that to handle it, I hope you won't be
- corpora can be your friend ;-)

FIGURE 35

analysis than is often done, because we don't always have multifactorial analysis, we don't always have nice random-effects structures, and as you've seen in this particular case, we don't always have nice controls for learning or habituation during an experiment that would need to be controlled. Actually, if you look at a lot of psycholinguistic literature, you don't even get  $R^2$ -values. What that means is a lot of the time, the statistical analysis that is reported, you don't even know at the end how good it is. Like, does it explain ten percent of the variability? Fifty percent? Ninety percent? Most of the time that doesn't get said, and so these are kinds of things we do want to control for.

Corpus studies likewise need better statistical analysis than is often employed. If you have an experimental design that, for instance, is completely based on a certain direction of association, like in this case, verb-to-construction, the sentence fragment the subjects were asked to complete ended in a verb with an instruction to complete it with some construction, then that's the kind of association measure that should be used. These kinds of things are really important in particular also then with keeping association and frequency separate. Alright, I'm going to skip this one.

I hope to have shown during the course of this week that corpus data have a lot to offer. Yes, you can work with frequencies, but really, please don't stop at that—use all these other things that we have been talking about, especially given the high degree of ecological validity that corpus data have to offer. But, you know, with this richness of information that you have in corpus data and the noisiness and the skewed distributions of this kind of data, comes a

responsibility, namely to understand what corpora can do and what they cannot do. And yes, I'm sorry, you need some degree of statistical expertise, because you need to define things properly, you need to analyze things properly, especially when you want to put claims to the test that come out of the usage-based kind of approach. Again, I hope that this kind of somewhat sobering or maybe scary conclusion that you do need a lot of quantitative methods does not scare you off.

Thank you.

# References

- Adelman, James S., Gordon D. A. Brown, and José F. Quesada. 2006. Contextual diversity, nor word frequency, determines word-naming and lexical decision times. *Psychological Science* 17(9), 814–823.
- Ambridge, Ben, Anne Theakston, Elena V. M. Lieven, and Michael Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract grammatical construction. *Cognitive Development* 21(2), 174–193.
- Anderson, John R. 2009. *Cognitive Psychology and Its Implications* (7th ed). New York: Worth.
- Aslin, Richard N. and Elissa L. Newport. 2012. Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science* 21(3), 170–176.
- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics with R*. Cambridge: Cambridge University Press.
- Baayen, R. Harald. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5(3), 436–461.
- Ball, Linden J., Jaswinder Shoker, and Jeremy N. V. Miles. 2010. Odour-based context reinstatement effects with indirect measures of memory: The curious case of rosemary. *British Journal of Psychology* 101(4), 655–678.
- Balota, David A. and Daniel H. Spieler. 1998. The utility of item level analyses in model evaluation: a reply to Seidenberg and Plaut. *Psychological Science* 9(3), 238–240.
- Barnwell, Brendan B. 2014. Effects of Nonlinguistic Context on Language Production. Unpublished Ph.D. dissertation.
- Bartlett, F.C. 1932 (1967). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.
- Beckner, Clay, Nick C. Ellis, Richard Blythe, John Holland, Joan Bybee, Jinyun Ke, Morten H. Christiansen, Diane Larsen-Freeman, William Croft, and Tom Schoenemann. 2009. Language is a complex adaptive system: position paper. *Language Learning* 59 (Suppl. 1), 1–26.
- Bernolet, Sarah, Timothy Coleman, and Robert Hartsuiker. 2014. The 'sense boost' to dative priming: evidence for sense-specific verb-structure links. *Journal of Memory and Language* 76(1), 113–126.
- Blumenthal-Dramé, Alice. 2016. What corpus-based Cognitive Linguistics can and cannot expect from neurolinguistics. *Cognitive Linguistics* 27(4), 493–505.
- Bock, J. Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3), 355–387.
- Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

- Boyd, Jeremy K. and Adele E. Goldberg. 2011. Learning what not to say: the role of statistical preemption and categorization in a-adjective production. *Language* 87(1), 55–83.
- Branigan, Holly P., Martin J. Pickering, Simon P. Livsledge, Andrew J. Stewart, and Thomas P. Urbach. 1995. Syntactic priming: investigating the mental representation of language. *Journal of Psycholinguistic Research* 24(6), 489–506.
- Bybee, Joan. 2006. From usage to grammar: the mind's response to repetition. *Language* 82(4), 711–733.
- Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.
- Bybee, Joan and Clay Beckner. 2009. Usage-based theory. In Bernd Heine and Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 827–856. Oxford: Oxford University Press.
- Bybee, Joan and Paul J. Hopper. 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam and Philadelphia: John Benjamins.
- Bybee, Joan and James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22(2–4), 381–410.
- Bybee, Joan and Sandra A. Thompson. 1997. Three frequency effects in syntax. *Berkeley Linguistics Society* 23, 65–85.
- Callies, Marcus. 2013. Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics* 18(3), 357–390.
- Casenhiser, Devin M. and Adele E. Goldberg. 2005. Fast mapping of a phrasal form and meaning. *Developmental Science* 8(6), 500–508.
- Cattell, James M. 1886. The time it takes to see and name objects. *Mind* x1, 63–65.
- Chang, Franklin., Gary S. Dell, J. Kathryn Bock, and Zenzi Griffin. 2000. Structural priming as implicit learning: a comparison of models of sentence production. *Journal of Psycholinguistic Research* 29(2), 217–229.
- Christiansen, Morten H. and Nick Chater. 2016. *Creating Language: Integrating Evolution, Acquisition, and Processing*. Cambridge, MA: The MIT Press.
- Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36(3), 181–204.
- Dąbrowska, Ewa. 2008. Words as constructions. In Vyvyan Evans and Stephanie S. Pourcel (eds.), *New Directions in Cognitive Linguistics*, 201–223. Amsterdam and Philadelphia: John Benjamins.
- Dąbrowska, Ewa. 2016. Cognitive Linguistics' seven deadly sins. *Cognitive Linguistics* 27(4), 479–491.
- Demberg, Vera and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2), 193–210.

- Divjak, Dagmar S., Natalia Levshina, and Jane Klavan. 2016. Cognitive linguistics: looking back, looking forward. *Cognitive Linguistics* 27(4), 447–463.
- Doğruöz, A. Seza and Stefan Th. Gries. 2014. Spread of on-going changes in an immigrant language: Turkish in the Netherlands. In Martin Pütz, Justyna Robinson, and Monika Reif (eds.), *Cognitive Sociolinguistics: Social and Cultural Variation on Cognition and Language Use*, 161–185. Amsterdam and Philadelphia: John Benjamins.
- Dunn, James. 2018. Multi-unit association measures: moving beyond pairs of words. *International Journal of Corpus Linguistics* 23(2), 183–215.
- Durham, Mercedes. 2011. I think (that) something's missing: Complementizer deletion in non-native emails. *Studies in Second Language Learning and Teaching* 1(3), 421–445.
- Ebbinghaus, Hermann. 1885. *Memory: A contribution to Experimental Psychology*. New York: Teachers College, Columbia University.
- Ellis, Nick C. 2002. Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition* 24(2), 143–188.
- Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1), 1–24.
- Ellis, Nick C. and Rita Simpson-Vlach. 2005. An academic formulas list (AFL): extraction, validation, prioritization. Paper presented at Phraseology 2005, Université Catholique Louvain-la-Neuve.
- Ellis, Nick C., Rita Simpson-Vlach and Carson Maynard. 2007. The processing of formulas in native and L2 speakers: psycholinguistic and corpus determinants. Paper presented at the Symposium on Formulaic Language, University of Wisconsin-Milwaukee.
- Ellis, Nick C., Ute Römer, and Matthew Brook O'Donnell. 2016. *Usage-based Approaches to Language Acquisition and Processing*. New York: Wiley-Blackwell.
- Fidelholtz, James L. 1975. Word frequency and vowel reduction in English. *Chicago Linguistic Society* 11, 200–213.
- Firth, John R. 1957. *Papers in Linguistics*. Oxford: Oxford University Press.
- Forster, Kenneth I. and Susan M. Chambers. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12(6), 627–635.
- Francom, Jerid. 2009. Experimental Syntax: Exploring the Effect of Repeated Exposure to Anomalous Syntactic Structure: Evidence from Rating and Reading Tasks. Ph.D. dissertation, University of Arizona.
- Gablasova, Dana., Vaclav Brezina, and Tony McEnery. 2017. Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning* 67(S1), 130–154.
- Geeraerts, Dirk. 2006. Methodology in cognitive linguistics. In Gitte Kristiansen, Michel Achard, René Dirven and Francisco J. Ruiz de Mendoza Ibáñez (eds.), *Cognitive*



- Linguistics: Current Applications and Future Perspectives*, 21–49. Berlin and New York: Mouton de Gruyter.
- Glenberg, Arthur M. 1976. Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior* 15(1), 1–15.
- Glenberg, Arthur M. 1979. Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory and Cognition* 7(2), 95–112.
- Godden, Duncan R. and Alan D Baddeley. 1975. Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology* 66(3), 325–331.
- Goldberg, Adele E., Devin M. Casenhiser, and Nitya Sethuraman. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 15(3), 289–316.
- Gómez, Rebecca L. 2002. Variability and detection of invariant structure. *Psychological Science* 13(5), 431–436.
- Gries, Stefan Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2), 109–151.
- Gries, Stefan Th. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In Stefan Th. Gries, Stefanie Wulff, and Mark Davies (eds.), *Corpus Linguistic Applications: Current Studies, New Directions*, 197–212. Amsterdam: Rodopi.
- Gries, Stefan Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language* 36(3), 477–510.
- Gries, Stefan Th. 2015. More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3), 505–536.
- Gries, Stefan Th. 2018. On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. *Journal of Second Language Studies* 1(2), 276–308.
- Gries, Stefan Th. 2019. 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics* 24(3), 385–412.
- Gries, Stefan Th. and Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. In Jesús Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*, 35–54. Cham: Springer.
- Gries, Stefan Th. and Sandra C. Deshors. 2014. Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora* 9(1), 109–136.
- Gries, Stefan Th. and Stefanie Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3, 182–200.

- Gries, Stefan Th. and Stefanie Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7, 163–186.
- Gries, Stefan Th. forthcoming. Priming of syntactic alternations by learners of English: an analysis of sentence-completion and collocation results. In Jesse A. Egbert and Paul Baker (eds.), *Using corpus methods to triangulate linguistic analysis*, 219–238. New York and London: Routledge.
- Halliday, Michael A. K. 2005. *Computational and Quantitative Studies*. London and New York: Continuum.
- Hardie, Andrew. 2008. message # 12240 to Corpora List, 14 August 2008.
- Harris, Zelig S. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel.
- Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
- Horton, William S. 2007. The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes* 22(7), 1114–1139.
- Howes, Davis H. and Richard L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41(6), 401–410.
- Huttenlocher, Janelle and Lorraine F. Kubicek. 1983. The source of relatedness effects on naming latency. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9(3), 486–496.
- Jaeger, T. Florian. 2010. Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology* 61(1), 23–62.
- Jaeger, T. Florian. 2011. Corpus-based research on language production: information density and reducible subject relatives. In Emily M. Bender and Jennifer E. Arnold. (eds.), *Language from a Cognitive Perspective: Grammar, Usage, and Processing. Studies in honor of Tom Wasow*, 161–197. Stanford: CSLI Publications.
- Jaeger, T. Florian and Neal Snider. 2008. Implicit learning and syntactic persistence: surprisal and cumulativity. *Proceedings of the Cognitive Science Society Conference*, 1061–1066.
- Jaeger, T. Florian and Neal E. Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition* 127(1), 57–83.
- Kaschak, Michael P., Timothy J. Kutta, and J. Leah Jones. 2011. Structural priming as implicit learning: cumulative priming effects and individual differences. *Psychonomic Bulletin and Review* 18(6), 1133–1139.
- Kelly, Michael H. 1986. On the Selection of Linguistic Options. Unpublished Ph.D. dissertation, Cornell University.
- Lachman, Roy. 1973. Uncertainty effects on time to access the internal lexicon. *Journal of Experimental Psychology* 99(2), 199–208.

- Lachman, Roy, Juliet Popper Shaffer, and Deborah Hennrikus. 1974. Language and cognition: effects of stimulus codability, name-word frequency, and age of acquisition on lexical reaction time. *Journal of Verbal Learning and Verbal Behavior* 13(6), 613–625.
- Lakoff, George. 1991. Cognitive versus generative linguistics: how commitments influence results. *Language and Communication* 11(1–2), 53–62.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar I: Theoretical Prerequisites*. Stanford: Stanford University Press.
- Langacker, Ronald W. 2016. Working toward a synthesis. *Cognitive Linguistics* 27(4), 465–477.
- Leech, Geoffrey N. 1992. Corpora and theories of linguistic performance. In Jan Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, Stockholm, 4–8 August, 105–122. Berlin and New York: Mouton de Gruyter.
- Lester, Nicholas A. 2019. *That's hard*: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research* 5(1), 1–32.
- Lester, Nicholas A., Daniel Baum, and Tirza Biron. forthcoming. Phonetic duration of nouns depends on de-lexicalized syntactic distributions: Evidence from naturally occurring conversation. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Lester, Nicholas A. and Fermín Moscoso del Prado Martín. 2017. Syntactic flexibility in the noun: evidence from picture naming. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 2585–2590.
- Lester, Nicholas A., Laurie B. Feldman, and Fermín Moscoso del Prado Martín. 2017. You can take a noun out of syntax ...: Syntactic similarity effects in lexical priming. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2537–2542.
- Linzen, Tal and T. Florian Jaeger. 2015. Uncertainty and expectation in sentence processing: evidence From subcategorization distributions. *Cognitive Science* 40(6), 1382–1411.
- McDonald, Scott A. and Richard C. Shillcock. 2001. Rethinking the word frequency effect: the neglected role of distributional information in lexical processing. *Language and Speech* 44(3), 295–323.
- McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Neset, Tore. 2016. Does historical linguistics need the cognitive commitment? Prosodic change in East Slavic. *Cognitive Linguistics* 27(4), 573–585.
- Newmeyer, Frederick J. 2006a. Grammar and usage: a response to Gregory R. Guy. *Language* 82(2), 399–404.

- Newmeyer, Frederick J. 2006b. On Gahl and Garnsey on grammar and usage. *Language* 82(4), 705–706.
- Oldfield, R. and A. Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology A*(17), 273–281.
- Onnis, Luca., Padraic Monaghan, Morten H. Christiansen, and Nick Chater. 2004. Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* 26, 1678–1683.
- Parker, Amanda., Henny Ngu, and Helen J. Cassaday. 2001. Odour and Proustian memory: Reduction of context-dependent forgetting and multiple forms of memory. *Applied Cognitive Psychology* 15(2), 159–171.
- Pickering, Martin J. and Holly P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language* 39(4), 633–651.
- Pickering, Martin J. and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36(4), 329–347.
- Rescorla, Robert A. 1988. Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience* 11(1), 329–352.
- Rescorla, Robert A. and Allen R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Abraham H. Black and William F. Prokasy (eds.), *Classical Conditioning II: Current Theory and Research*, 64–99. New York: Appleton-Century-Crofts.
- Rowland, Caroline F., Franklin Chang, Ben Ambridge, Julian M. Pine, and Elena V. M. Lieven. 2012. The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition* 125(1), 49–63.
- Rumelhart, David E., Geoffrey E. Hinton, Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature Reviews Neuroscience* 323(6088), 533–536.
- Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294), 1926–1928.
- Savage, Ceri., Elena V. M. Lieven, Anna Theakston, and Michael Tomasello. 2006. Structural priming as implicit learning in language acquisition: The persistence of lexical and structural priming in 4-year-olds. *Language Learning and Development* 2(1), 27–49.
- Scheepers, Christoph. 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition* 89(3), 179–205.
- Schmid, Hans-Jörg. 2010. Does frequency in the text instantiate entrenchment in the cognitive system? In Dylan Glynn and Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches*, 101–133. Berlin and Boston: Mouton de Gruyter.

- Schneider, Ulrike. forthcoming.  $\Delta P$  as a measure of collocation strength. *Corpus Linguistics and Linguistic Theory*.
- Schooler, Lael J., and John R. Anderson. 1997. The role of process in the rational analysis of memory. *Cognitive Psychology* 32(3), 219–250.
- Schuchardt, Hugo. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Oppenheim.
- Seidenberg, Mark S. and Mayellen C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive Science* 23(4), 569–588.
- Smith, Nathaniel and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128(3), 302–319.
- Smith, Steven M. 1979. Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory* 5(5), 460–471.
- Smith, Steven M. 1985. Background music and context-dependent memory. *The American Journal of Psychology* 98(4), 591–603.
- Smith, Steven M., Fred R. Heath, and Edward Vela. 1990. Environmental context-dependent homophone spelling. *The American Journal of Psychology* 103(2), 229–242.
- Snider, Neal. 2009. Similarity and structural priming. *Proceedings of the 31st Annual Conference of the Cognitive Science*, 815–820.
- Spieler, Daniel H. and David A. Balota. 1997. Bringing computational models of word naming down to the item level. *Psychological Science* 8, 411–416.
- Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair. In Mona Baker, F. Francis, and Elena Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, 1–46. Amsterdam and Philadelphia: John Benjamins.
- Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1), 113–150.
- Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English. A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis*. Berlin and New York: Mouton de Gruyter.
- Tagliamonte, Sali A. and Jennifer Smith. 2005. No momentary fancy! The zero ‘complementizer’ in English dialects. *English Language and Linguistics* 9(2), 289–309.
- Teubert, Wolfgang. 2005. My version of Corpus Linguistics. *International Journal of Corpus Linguistics* 10(1), 1–13.
- Thompson, Sandra A. and Anthony Mulac. 1991. The discourse conditions for the use of the complementizer that in conversational English. *Journal of Pragmatics* 15(3), 237–251.
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.

- Torres Cacoullos, Rena and James A. Walker. 2009. On the persistence of grammar in discourse formulas: A variationist study of that. *Linguistics* 47(1), 1–43.
- Trautscholdt, Martin. Experimentelle Untersuchungen über die Association der Vorstellungen. Unpublished Ph.D. dissertation, University of Leipzig.
- Wills, Andy J. 2009. Prediction errors and attention in the presence and absence of feedback. *Current Directions in Psychological Science* 18, 95–100.
- Wulff, Stefanie. 2016. A friendly conspiracy of input, L1, and processing demands: That-variation in German and Spanish learner language. In Lourdes Ortega, Andrea E. Tyler, Hae In Park, and Mariko Uno (eds.), *The Usage-based Study of Language Learning and Multilingualism*, 115–136. Georgetown: Georgetown University Press.
- Wulff, Stefanie, Nicholas Lester, and Maria T. Martinez-Garcia. 2014. That-variation in German and Spanish L2 English. *Language and Cognition* 6, 271–299.
- Wulff, Stefanie, Stefan Th. Gries, and Nicholas Lester. 2018. Optional that in complementation by German and Spanish learners. In Andrea Tyler and Carol Moder (eds.), *What is Applied Cognitive Linguistics? Answers from Current SLA Research*, 99–120. New York: Mouton de Gruyter.
- Wulff, Stefanie and Stefan Th. Gries. forthcoming. Explaining individual variation in learner corpus research: Some methodological suggestions. In Bert Le Bruyn and Magali Paquot (eds.), *Learner Corpora and Second Language Acquisition Research*. Cambridge: Cambridge University Press.

## About the Series Editor

Fuyin (Thomas) Li (1963, Ph.D. 2002) received his Ph.D. in English Linguistics and Applied Linguistics from the Chinese University of Hong Kong. He is professor of linguistics at Beihang University, where he organizes *China International Forum on Cognitive Linguistics* since 2004, <http://cifcl.buaa.edu.cn/Intro.htm>. As the founding editor of the journal *Cognitive Semantics*, [brill.com/cose](http://brill.com/cose), the founding editor of *International Journal of Cognitive Linguistics*, editor of the series *Distinguished Lectures in Cognitive Linguistics*, [brill.com/dlcl](http://brill.com/dlcl), (originally *Eminent Linguists' Lecture Series*), editor of *Compendium of Cognitive Linguistics Research*, and organizer of ICLC-11, he plays an active role in the international expansion of Cognitive Linguistics.

His main research interests involve the Talmyan cognitive semantics, overlapping systems model, event grammar, causality, etc., with a focus on synchronic and diachronic perspective on Chinese data, and a strong commitment to usage-based model and corpus method.

His representative publications include the following: *Metaphor, Image, and Image Schemas in Second Language Pedagogy* (2009), *Semantics: A Course Book* (1999), *An Introduction to Cognitive Linguistics* (in Chinese, 2008), *Semantics: An Introduction* (in Chinese, 2007), *Toward a Cognitive Semantics, Volume I: Concept Structuring Systems* (Chinese version, 2017), *Toward a Cognitive Semantics, Volume II: Typology and Process in Concept Structuring* (Chinese version, 2019).

His personal homepage: <http://shi.buaa.edu.cn/thomasli>

E-mail: [thomasli@buaa.edu.cn](mailto:thomasli@buaa.edu.cn); [thomaslfy@gmail.com](mailto:thomaslfy@gmail.com)

# Websites for Cognitive Linguistics and CIFCL Speakers

All the websites were checked for validity on 20 January 2019

## Part 1 Websites for Cognitive Linguistics

1. <http://www.cogling.org/>  
Website for the International Cognitive Linguistics Association (ICLA)
2. <http://www.cognitivelinguistics.org/en/journal>  
Website for the journal edited by ICLA, *Cognitive Linguistics*
3. <http://cifcl.buaa.edu.cn/>  
Website for China International Forum on Cognitive Linguistics (CIFCL)
4. <http://cosebrill.edmgr.com/>  
Website for the journal *Cognitive Semantics* (ISSN 2352-6408/ E-ISSN 2352-6416), edited by CIFCL
5. <http://www.degruyter.com/view/serial/16078?rskey=fw6Q2O&result=1&q=CLR>  
Website for the Cognitive Linguistics Research (CLR)
6. <http://www.degruyter.com/view/serial/20568?rskey=dddL3r&result=1&q=ACL>  
Website for Application of Cognitive Linguistics (ACL)
7. <http://www.benamins.com/#catalog/books/clsc/main>  
Website for book series in Cognitive Linguistics by Benamins
8. <http://www.brill.com/dlcl>  
Website for Distinguished Lectures in Cognitive Linguistics (DLCL)
9. <http://refworks.reference-global.com/>  
Website for online resources for Cognitive Linguistics Bibliography
10. <http://benamins.com/online/met/>  
Website for Bibliography of Metaphor and Metonymy



11. <http://linguistics.berkeley.edu/research/cognitive/>  
Website for Cognitive Program in Berkeley
12. <https://framenet.icsi.berkeley.edu/fndrupal/>  
Website for Framenet
13. <http://www.mpi.nl/>  
Website for the Max Planck Institute for Psycholinguistics

## Part 2 Websites for CIFCL Speakers and Their Research

14. CIFCL Organizer  
**Thomas Li**, [thomasli@buaa.edu.cn](mailto:thomasli@buaa.edu.cn); [thomaslfy@gmail.com](mailto:thomaslfy@gmail.com)  
Personal homepage: <http://shi.buaa.edu.cn/thomasli>  
<http://shi.buaa.edu.cn/lifuyin/en/index.htm>
15. CIFCL 18, 2018  
**Arie Verhagen**, [A.Verhagen@hum.leidenuniv.nl](mailto:A.Verhagen@hum.leidenuniv.nl)  
<http://www.arieverhagen.nl/>
16. CIFCL 18, 2018 (CIFCL 12, 2013)  
**Stefan Th. Gries**, [stgries@linguistics.ucsb.edu](mailto:stgries@linguistics.ucsb.edu)  
<http://www.stgries.info>
17. CIFCL 17, 2017  
**Jeffrey M. Zacks**, [jzacks@wustl.edu](mailto:jzacks@wustl.edu)  
Lab: [dcl.wustl.edu](http://dcl.wustl.edu)  
Personal homepage: <https://dcl.wustl.edu/affiliates/jeff-zacks/>
18. CIFCL 16, 2016  
**Cliff Goddard**, [c.goddard@griffith.edu.au](mailto:c.goddard@griffith.edu.au)  
<https://www.griffith.edu.au/humanities-languages/school-humanities-languages-social-science/research/natural-semantic-metalanguage-homepage>
19. CIFCL 15, 2016  
**Nikolas Gisborne**, [n.gisborne@ed.ac.uk](mailto:n.gisborne@ed.ac.uk)
20. CIFCL 14, 2014  
**Phillip Wolff**, [pwolff@emory.edu](mailto:pwolff@emory.edu)

21. CIFCL 13, 2013 (CIFCL 3, 2006)  
**Ronald W. Langacker**, [rlangacker@ucsd.edu](mailto:rlangacker@ucsd.edu)  
<http://idiom.ucsd.edu/~rwl/>
22. CIFCL 12, 2013 (CIFCL 18, 2018)  
**Stefan Th. Gries**, [stgries@linguistics.ucsb.edu](mailto:stgries@linguistics.ucsb.edu)  
<http://www.stgries.info>
23. CIFCL 12, 2013  
**Alan Cienki**, [a.cienki@vu.nl](mailto:a.cienki@vu.nl)  
<https://research.vu.nl/en/persons/alan-cienki>
24. CIFCL 11, 2012  
**Sherman Wilcox**, [wilcox@unm.edu](mailto:wilcox@unm.edu)  
<http://www.unm.edu/~wilcox>
25. CIFCL 10, 2012  
**Jürgen Bohnemeyer**, [jb77@buffalo.edu](mailto:jb77@buffalo.edu)  
Personal homepage: <http://www.acsu.buffalo.edu/~jb77/>  
The CAL blog: <https://causalityacrosslanguages.wordpress.com/>  
The blog of the UB Semantic Typology Lab: <https://ubstlab.wordpress.com/>
26. CIFCL 09, 2011  
**Laura A. Janda**, [laura.janda@uit.no](mailto:laura.janda@uit.no)  
<http://ansatte.uit.no/laura.janda/>  
[https://uit.no/om/enhet/ansatte/person?p\\_document\\_id=41561&p\\_dimension\\_id=210121](https://uit.no/om/enhet/ansatte/person?p_document_id=41561&p_dimension_id=210121)
27. CIFCL 09, 2011  
**Ewa Dąbrowska**, [ewa.dabrowska@northumbria.ac.uk](mailto:ewa.dabrowska@northumbria.ac.uk)
28. CIFCL 08, 2010  
**William Croft**, [wcroft@unm.edu](mailto:wcroft@unm.edu)  
<http://www.unm.edu/~wcroft>
29. CIFCL 08, 2010  
**Zoltán Kövecses**, [kovecses.zoltan@btk.elte.hu](mailto:kovecses.zoltan@btk.elte.hu)
30. CIFCL 08, 2010  
**(Melissa Bowerman: 1942–2011)**

31. CIFCL 07, 2009  
**Dirk Geeraerts**, [dirk.geeraerts@arts.kuleuven.be](mailto:dirk.geeraerts@arts.kuleuven.be)  
<http://wwwling.arts.kuleuven.be/qlvl/dirkg.htm>
32. CIFCL 07, 2009  
**Mark Turner**, [mark.turner@case.edu](mailto:mark.turner@case.edu)
33. CIFCL 06, 2008  
**Chris Sinha**, [chris.sinha@ling.lu.se](mailto:chris.sinha@ling.lu.se)
34. CIFCL 05, 2008  
**Gilles Fauconnier**, [faucon@cogsci.ucsd.edu](mailto:faucon@cogsci.ucsd.edu)
35. CIFCL 04, 2007  
**Leonard Talmy**, [talmy@buffalo.edu](mailto:talmy@buffalo.edu)  
<https://www.acsu.buffalo.edu/~talmy/talmy.html>
36. CIFCL 03, 2006 (CIFCL 13, 2013)  
**Ronald W. Langacker**, [rlangacker@ucsd.edu](mailto:rlangacker@ucsd.edu)  
<http://idiom.ucsd.edu/~rwl/>
37. CIFCL 02, 2005  
**John Taylor**, [john.taylor65@xtra.co.nz](mailto:john.taylor65@xtra.co.nz)  
<https://independent.academia.edu/JohnRTaylor>
38. CIFCL 01, 2004  
**George Lakoff**, [lakoff@berkeley.edu](mailto:lakoff@berkeley.edu)  
<http://georgelakoff.com/>