# Ten Lectures on Quantitative Approaches in Cognitive Linguistics

# Distinguished Lectures in Cognitive Linguistics

*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*

# Ten Lectures on Quantitative Approaches in Cognitive Linguistics

*Corpus-linguistic, Experimental, and Statistical Applications*

*By*

Stefan Th. Gries

BRILL

LEIDEN | BOSTON

The Library of Congress Cataloging-in-Publication Data is available online at http://catalog.loc.gov
LC record available at http://lccn.loc.gov/

~~Want or need Open Access? Brill Open offers you the choice to make your research freely accessible online in exchange for a publication charge. Review your various options on brill.com/brill-open~~.

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

ISSN 2468-4872
ISBN 978-90-04-33621-6 (hardback)
ISBN 978-90-04-33622-3 (e-book)

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

# Contents

# Preface

The present text, entitled *Ten Lectures on Quantitative Approaches in Cognitive Linguistics: Corpus-linguistic, experimental, and statistical applications* by Stefan Gries is a transcribed version of the lectures given by Professor Stefan Gries in May 2013 as the forum speaker for *the 12th China International Forum on Cognitive Linguistics.* Stefan 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) and Honorary Liebig-Professor 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 test and develop corpus-linguistic and statistical methods. Theoretically, he is a cognitively-oriented usage-based linguist. Gries has comprehensive publications in books, co-edited volumes, and articles in the leading peer-reviewed journals of his fields (Cognitive Linguistics and International Journal of Corpus Linguistics) and many other peer-reviewed journals. He is founding editor-in-chief of the international peer-reviewed journal *Corpus Linguistics and Linguistic Theory*, associate editor of *Cognitive Linguistics*, and performs editorial functions for many international peer-reviewed journals and book series. The following are just a few of his most representative publications: *Statistics for linguistics with R* (2013); *Quantitative corpus linguistics with R: a practical introduction* (2009); The genitive alternation in Chinese and German ESL learners: towards a multifactorial notion of context in learner corpus research (2013); Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications (2012); and Dispersions and adjusted frequencies in corpora (2008). For more information, pls visit his website at: http://www.linguistics.ucsb.edu/faculty/stgries/index.html.

*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. It is a continuing program organized by several prestigious universities in Beijing. The following is a list of organizers for CIFCL 12.

*Main organizer*
Li Fuyin (Thomas), PhD/Professor, Beihang University

*Co-organizers*
Liu Shisheng, PhD/Professor, Tsinghua University
Gao Yihong, PhD/Professor, Peking University
Shi Baohui, PhD/Professor, Beijing Forestry University
Zhang Xu, PhD/Associate Professor, Beijing Language University

Professor Stefan Gries*’s* lecture series was mainly supported by *the Beihang Grant for International Outstanding Scientists* for 2013 (Project number: Z1359, Project organizer: Thomas Fuyin Li).

The text is published, accompanied by its videodisc counterpart and Chinese guide, as one of the *Eminent Linguists Lecture Series*. The transcription of the video, proofreading of the text, writing the Chinese guide, and publication of the work in its present book form, have involved many people’s strenuous inputs. The initial transcripts were completed by the following postgraduate students: Du Jing, Liu Yunfeng, Wu Xiaoqing, Li Heng, Yu Lin, Ren Longbo, Liu Jia, Deng Yu, He Yuanyuan. Deng Yu, Jia Hongxia, and Weiqing had revisions for the whole text. Then we editors did the 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: […]. The stage direction describes what the speaker was doing, such as pointing at a slide, showing an object, etc. The speaker, Professor Stefan Gries did the final revisions. The published version is the final version approved by the speaker.

The publication of this book is sponsored by the National Social Science Foundation Award No.13BYY012, and a Humanities and Social Science Program Award from the Ministry of Education No.09YJA740010.

*Thomas Fuyin Li*
Beihang University (BUAA)

*Yan Ding*
Beijing Jiaotong University

*Weiwei Zhang*
Shanghai Foreign Studies University

# About the Author

Stefan Gries is Professor of Linguistics in the Department of Linguistics at the University of California, Santa Barbara. As a quantitative corpus linguist at the intersection of corpus linguistics, cognitive linguistics, and computational linguistics, Professor Gries uses a range of statistical methods to investigate various topics in morpho-phonology, syntax, semantics, as well as first and second/foreign language acquisition. Professor Gries has comprehensive publications including books, co-edited volumes, and numerous articles in the leading journals of his fields, viz. *Cognitive Linguistics* and *International Journal of Corpus Linguistics*, and many other international peer-reviewed journals. He is founding editor-in-chief of the journal *Corpus Linguistics and Linguistic Theory*, associate editor of *Cognitive Linguistics* and *Cognitive Linguistic Studies*. He also performs editorial functions for the book series *Cognitive Linguistics and Practice, Corpora and Language in Use*, and *Explorations in English Language and Linguistics*.

LECTURE 1

# Corpus Linguistics, Cognitive Linguistics, and Psycholinguistics: On their Combination and Fit

Thank you very much for the invitation and I'm very happy and honored to be here. I will talk a little bit about some of the things I have been doing in the past. The talks that I will give are all sort of talks about some different types of corpus linguistics work, sometimes in combination with other corpus linguistic work, sometimes with experimental work, sometimes with a lot of different types of statistical applications. The talks will build on each other in some sense, it doesn't mean that one has to listen to them in sequence, but if you were to listen tomorrow, then you would see how some later talks will come back to some of the things I talked about at the beginning. So today's talk will be more theoretical in nature and less empirical. I want to talk a little bit about, as the title suggests, about how I think that corpus linguistics, cognitive linguistics and psycholinguistics fit together because if you look at some of the discussions that have been dominating, especially, the recent field of corpus linguistics, then I think it's important to point out some commonalities and some shared assumptions, and some terminology, and then a lot of the later talks will basically try to come back to many of these issues and revisit them with what I hope are also state-of-the-art applications.

So, as I say here, in general, if you look at corpus linguistics and linguistic theory, so not even cognitive linguistics per se but linguistics theory in general, then that relationship has been problematic, to say the least in some sense. Because on the one hand there are a lot of different opinions from corpus linguists on what corpus linguistics actually is, and here are some of the terms that have been thrown out to talk about corpus linguistics. Some people say it's a tool or it's a method, or an approach or a discipline and all these other notions. Then, there are some ways in which corpus linguists in particular have talked about the field that actually make it relatively unappealing to people who have more theoretical or cognitive interests, which is of course regrettable. To show you some examples of these and to mention some people who have talked about things like that, some people who have said it's a theory would be, for instance, Geoff Leech, who said "Computer corpus linguistics defines not just newly emerging methodology but in fact a new research enterprise or philosophical approach." Stubbs has said "Corpus is an important

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_002

concept in linguistic theory or is something that possesses theoretical status," all sorts of opinions that gone on in that direction. Particularly influential, at this point, because she is the editor of the leading corpus linguistic journal, would be Michaela Mahlberg, who said it's a corpus-theoretical approach. Then Wolfgang Teubert, the previous editor of that same most influential journal, who would say it's a theoretical approach. Now other people say the exact opposite in a way, namely, that it's *only* in air quotes—a methodology. McEnery and Wilson's very ell-known textbook and Meyer's well-known textbook, Bowker and Pearson, from the domain of applied linguistics, they all basically just say "well, it's a method and not a theory in and of itself". McEnery, in a more recent textbook, put it even more concisely—with colleagues—as a whole, it's a system of methods and principles and it has a theoretical status but that doesn't make it a theory in and of itself. And a recent posting on the LinguistList, Andrew Hardie would say something like this, namely, as a corpus linguist, he considers himself primarily a methodologist and then the method can be applied to whatever theory you find interesting. Like I said, there is a variety of other labels out there like *discipline* mentioned by these people or *a methodological commitment*. Now the thing is when people talk about these things, usually you can locate them on one continuum, on another continuum. So this question of whether it is a theory or not is often correlated with what that people consider themselves, what I would call *corpus-driven* or *corpus-based* corpus linguists. Now what does that mean? Corpus-driven, the idea is that you build linguistic theory from scratch. That is, you look at your corpus data without any prior theoretical commitments. Nothing that you talk about does not come from the corpus. Here is an example, an exhaustive quote: While corpus linguistics may make use of the categories of traditional linguistics? It does not take them for granted. It's the discourse or the corpus itself, and not a language-external taxonomy that provides categories, classifications, theories, hypotheses and so on. Teubert actually went so far as to say, which I, as you will see in a moment, completely disagree with, "Corpus linguists still don't know what a morpheme, a phrase or a pattern is." Now corpus-based linguists, on the other hand, are sort of less radical in that sense, they approach corpus data from the perspective of some moderate corpus-external premises, and the idea is to take the theory or hypotheses that you have and test them in the face of a type of corpus data that you might have, which also means that this type of approach usually makes heavy use of annotation, either existing annotation in corpora or an annotation that is done on the basis of some corpora. Now what do I think? I do think that corpus linguistics is a major methodological paradigm, but I do not consider it a theory, if anything I would say methodological commitment. And why is that? Basically for two reasons. One is that I kind of feel uncomfortable with the idea that you have a theory whose name is a source of data. Second, I have even bigger problems with the idea that corpus-driven perspective per se is a theory. Now let me elaborate briefly on those two things. One, again, was the thing, a theory that is the name of a source of data. Even one of the people who helped coin the term *corpus linguistics* is reported as having had some awkward feelings about this because again—this is how it's said here—that it's not very good idea to have a discipline be called by the major research tool and data source but not by what the discipline maybe actually tries to achieve or what the theoretical commitments are. And so in a way, and maybe somewhat more polemically speaking, I wouldn't say corpus linguistics is a theory because I also don't say there is a theory called experimental linguistics or eye-tracking linguistics or self-paced reading linguistics. All of these are methods and all these are good methods and methods that exist for very real purposes but that doesn't make them any theory per se.

Now what about the second reason why I don't think corpus linguistics is a theory, the association with corpus-driven linguistics? There is a lot of papers and studies out there that say that they are corpus-driven, but actually I think most of them are not. I think I still have yet to see a really corpus-driven approach. And I want to show a few examples of how this is true. Some of them have to do with the notions of lexis and grammar and a lot of times lexical semantics. And the idea here is, if you really approach a corpus without any prior commitments as to what is in the corpus, then that actually means, that would go so far as to say that you don't know what a word is or what a morpheme is or what a syntactic pattern is. I mean that's because a word that's already a theoretical notion, or a morpheme is a theoretical notion. So if you really were corpus-driven then basically you, the first section of any your papers, would have to say "well I think *this* is word and *this* is a morpheme". Of course, no one does that. So of course even the most corpus-driven linguists that are out there, the most ardent defenders of that theory, I mean, never jump through these hoops. Like I said before, even if Teubert says "We still don't know what a morpheme, a word, a phrase or something like that". Then one of his closest colleagues/collaborators, Bill Louw, has no problems annotating or doing a concordance of *all sorts of* without having a whole section in the paper that says "that is actually word and that is how I know it."

So a lot of times, corpus-driven studies are really not corpus-driven at all. And one of the main people who is often cited in defense of corpus-driven linguistics has actually said exactly that. Halliday wrote "a corpus-driven grammar is not one that is theory-free," because you always bring some prior conceptions to the data. The only idea would be that hopefully you can validate them. And in a very nice overview article, in Mouton's recent *Handbook on Corpus Linguistics*, Richard Xiao wrote "applying intuitions from classifying concordances might simply be an implicit annotation process, which unconsciously makes use of preconceived theory, and this implicit annotation is to all intents and purposes unrecoverable and thus more unreliable than explicit annotation."

So the idea is, if you say you are corpus-driven but still use some pretheoretical notions, you basically get around to really rigorously identifying what you are concerned with, and thus, in a way, you make your study even more attackable than if it was corpus-based. Now here are some more practical examples, so less theoretical ones. So, for instance, there's a lot of corpus-driven studies right now that look at *n*-grams, where *n*-grams is basically defined as 'a sequence of *n* words.' So *in spite of* would be a 3-gram, *because of* could be a 2-gram or something like that, any sequences of several words, usually contiguous. And right now there are a lot of people who look at 4-grams for some reason. Unfortunately, I think the main reason why they do that is that it generates a lot of results but not too many, so you can still handle them. But the 4, I mean the number 4 per se, is not theoretically motivated in any way. But what people don't do, although they say they are corpus-driven, is show that 4 is actually the most useful number to look at. Because there's a lot of cases of course where 4 would not be the best number. So here are some examples, so if you look at 4-grams because that's hip right now. Then of course you miss *of course* because for that you would need it to be two. You miss *in spite of* because for that it would be 3, *on the one hand*, well yeah that will be actually a 4-gram. But then *as a matter of fact*, you will not get that because that's a 5-gram and so on. So basically what you will really need for a truly corpus-driven approach would be a number *n* that can vary, depending on your needs and depending on the type of things that you want to look at. But right now even in corpus-driven linguistics supposedly, most people don't even do that. Let me skip that. Actually there's a lot of work out there, especially in the last few years, that tried to come up with algorithms, programming ways to basically find what the best number is, or what the best approach to define *n* is. I am not going to detail about this right now. If you have question about this later, feel free to ask me of course.

Final type of argument about corpus-driven linguistics: If you think about it in terms of register, a situationally-defined text type, for instance, genres like newspaper language versus academic writing versus spoken conversation and things like that, then also a lot of types of corpus-driven work is actually not corpus-driven in any way. Again one of the favorite people quoted by corpus-driven linguists said this: "Register variation can in fact be defined as systematic variation in probabilities, in probabilities of occurrences of words, of constructions, and or of words in constructions. A register or genre is a tendency to select certain combinations of meanings with certain frequencies." Now, if you are a corpus-driven linguist, that basically means that any distinction of a corpus or division of corpus into registers, you should not take it for granted, because, mostly, if something corpus compilers, who collected the data, came up with for sampling reasons or maybe even for convenient reasons. But that does not mean they have any theoretical relevance. So from a truly bottom-up perspective, the distinction between speaking and writing might not be particularly meaningful or insightful, or the distinction between written-to-be-printed and written-to-be-spoken, that is a distinction that corpus compilers often make, but in a truly bottom-up perspective it might not be even relevant. So for example, if you look at how strongly 2-grams are attracted to each other, so things like *because of, according to, of course*—how strong are these two words attracted to each other?—then you find that a distinction that corpus linguists love to make, namely, for instance between academic writing and newspaper language, is hardly different at all. You find the same types of 2-grams with very similar degrees of attractions, so you wouldn't actually have to make that distinction. Same thing here, corpus linguists most of the time make a really big deal out of speaking versus writing, that the two are so different and that speaking is so much diverse in everything. But for some phenomena that might not be the case. I am not saying in general it's important. I am also not saying in general it's unimportant. I am saying it's just we have to look and we can't take anything for granted. Now if you look at the ditransitive construction, something like *he gave him the book, he sent him a present* or something like that, then the types of verbs that like to go into this construction, they don't differ at all or hardly in terms of speaking versus writing. So it doesn't matter whether you make that distinction or not. So a lot of times people sort of take these distinctions for granted even if they adopt corpus-driven approach although they wouldn't have to.

So to sum up that part, and again Xiao's overview article does it in a very beautiful and explicit way although, as you will see in a moment, I disagree with one central direction of his argument. So he says the distinction between corpus-driven, supposedly atheoretical, and corpus-based work is "overstated". Well, I will get back to that. And secondly he says the corpus-based approach is actually better suited to contributing to linguistic theory. Now I think that if anything the distinction is *under*stated, given that I think actually truly corpus-driven work is really hard to find. I've been along for a number of years; I think maybe there are a handful of articles that are truly corpus-driven, and the rest is not. If anything I think, No, we should make very clear that these two things are completely different and that one of them actually doesn't exist in the way that people say it exists. And second what that means is that indeed—I think this is where Richard is right—that I think corpus-driven linguistics which uses corpus-driven characteristics to say "corpus linguistics is a theory" is in fact less suited to contributing to corpus linguistic theory because many of the assumptions that it uses in the analysis of data are either unwarranted or are not even backed up or topicalized.

Now, what about the relationship between corpus linguistics and linguistic theory in general? And here I have to get a little bit personal because some people started it. So, some corpus linguists are not, as I say here, are not concerned with linguistic system that theoretical linguists may care about. So as a theoretical linguist or as a corpus linguist we a lot of time try to, I mean, we look at linguistic data but what we want to find out is basically how the mind works. I mean how the mind processes language. If we talk about language acquisition we want to figure out how it is possible that the linguistic system of a child, as she or he grows up can learn all this stuff so quickly, which is an essentially psycholinguistic or a cognitive question. Many corpus linguists couldn't care less, which means they might use corpora for practical or applying purposes like lexicography and language teaching, and of course that's not bad—it's just a very different emphasis. They might not be interested in linguistic theories at all, and again, that is fine. That's of course their prerogative. But sometimes they also have really, what I in a friendly way want to call, unusual ideas about potentially irrelevant neighboring disciplines. They have unusual ways of defending their perspectives. And finally they have unusual ideas about the nature of the discipline. By discipline, I mean linguistics and corpus linguistics that go beyond these issues here. So let me give you some examples of unusual ideas about potentially relevant neighboring disciplines. In a programmatic paper Teubert at some point wrote this "on the relationship between cognitive linguistics—obviously our topic here—and natural language processing, which I mean it's not the same as corpus linguistics, I think, but at least related, he said "the latter—corpus or computational linguistics—is the illegitimate offspring of the former." I don't even know what that means but that's what he thought. That's why I think it's unusual. I mean I don't even see the relation between the two. Here is a second example. It's kind of a long quote, and we don't have to go through this in detail but I want to show you this one example for some very strange views of corpus linguistics on what happens in cognition and in psycholinguistics. So Mason, who has actually done really interesting work in other areas, he says "formal approaches like generative grammar take for granted a hierarchical phrase structure. However, language is not produced in that way. But instead, it is produced as a linear sequence created in stops and starts." Ok, I mean at least if you look at the output of the speaker then that is probably true. But then he says "the hierarchical structure that cannot account for the fact at the beginning of the utterance is already produced before the whole sentence has been completely worked out. Similar issues apply for the reception of language." In general, that is true. We begin producing a sentence, I mean we utter the first words when we might not have been finishing with completing the process of the sentences in what we will say. So we start talking verbally before our syntactic planning has been completed. However, that does not mean that whatever planning processes go into this are not hierarchical. Just because the output sort of goes out from left to right or from early in time to later in time, sort of looks linear, doesn't meant that production processes behind it are [not] completely hierarchical. I mean there is a lot of good evidence for some sort of hierarchical phrase structure. But some people in corpus linguistics basically just gloss over that fact, and that of course makes it much more difficult for them to relate to cognitive linguistics, and to psycholinguistics than it would have to be.

Now some other ways in which corpus linguistics have disagreed with linguistic theory are a little bit more political in nature, unfortunately. So for some people, there is a very strong sentiment of 'we have to defend our pure field of corpus linguistics against what other people do' with some sort of warfare rhetoric. So corpus linguistics or particular types of corpus linguistic approaches have basically been discredited as not being British corpus linguistics, which obviously is not that helpful. Or because they do damage to good Sinclairian type of corpus linguistics which again doesn't really help. "The label *corpus linguistics* has been hijacked by theoretical linguists of all feathers." Again note that this is actually really only a problem if you think corpus linguistics is a theory. If you think it's a method then you will be grateful. Right, I mean how great it is that all these people do these different theories and they all use corpus data? But of course if you think it's a theory then of course you don't want people who have other theories to take over your theory, a little counterintuitive in a way. And that even went so far as to lead some people, again, like Wolfgang Teubert—I don't have a personal thing with him; it's just he has been very vocal about this as the editor of the *International Journal of Corpus Linguistics*—and he has even argued against the use of some software that people use for corpus linguistics because he says "it doesn't matter what kind of strings of information be our processes. It could be language but also it could be DNA sequences, or the numbers of *pi*" or something like that. Again I don't think that makes a lot of sense because any corpus software, I mean, you can apply to any other type of data. Of course you can apply corpus software to corpus but you can also apply to your shopping list. That doesn't make it any less of a relevant corpus linguistic tool.

Finally, some rather unusual ideas about the nature of the discipline, again to remind you I mean here, corpus linguists having weird ideas about what corpus linguistics do but also what corpus linguists do. For instance, I was told at some point, that corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions. I beg to differ. I know a lot of corpus linguists in general who are really interested in general rules and assumptions and want to find out how language works in general and not just in a particular text. So it was then recommended that when linguists come across a sentence such as "The sweetness of this lemon is sublime", then the task of the linguist should "be to look to see if other testimony in the discourse does or does not provide supporting evidence." I don't know whether that really would be a linguistic question necessarily. I do think, though, that corpus linguists might look at the sentence with interesting general principles or general rules, some syntactic structure or something like that without having to look for, there is some else who also find that the lemon was sweet, again, it's just a very narrow view on focusing on discourse. People have said "corpus linguistics looks at language from a social perspective but is not concerned with the psychological aspects of language", which is an empirical statement might or might not be true. But then on the other hand, it has been argued "linguistics is not a science"—by the same people—"like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of the endeavor to make sense of the human condition." Now it is not quite clear to me how you can be interested in human condition but blank out psychological aspects of language. Because obviously, when we use language, a lot of our psychology, a lot of our cognitive processing aspects and abilities feature in this.

So to sum up, corpus linguistics is in a way a very young discipline, and I think a lot of its work has left a mark on theoretical linguistics, psycholinguistics, applied linguistics. But corpus linguistics in general should interact a lot more with people from other neighboring disciplines. I think, as you will see in a moment, that cognitive linguistics would be a prime candidate for this type of interaction. And this is because a lot of corpus linguists have taken some sort of political stances or delimitations of the fields and have taken it for granted and don't validate them against what's happening in neighboring fields that have a lot to offer. One example, for instance, would be measures of dispersion. Measures of dispersion are statistics that indicate how widely words, for instance, constructions are distributed in a corpus. So if you have a very large corpus and a particular word might show up a hundred times in it and it might show up like everywhere, randomly distributed through the corpus. Or you might have the same type of corpus, another word, showing up a hundred of times there but only in one or two of the files, so it would be a very specialized word. So things like that, as I will argue in a moment, are things that will be really relevant to look at for corpus linguistics because of the cognitive implications. Same for measures of collocational strength. Collocational strength refers to the fact that, for instance, one word might be very strongly attracted to another word so there is a high tendency for the two words to co-occur together. So something like *according to*, for instance, would be an example, if you take any English text and you see the word *according*, there is really a high chance the next word will be *to*. I mean there is hardly anything else can come after it. Another example would be, if you read an English text and you will find *course*, chances are really high that the word before that will be *of*, because of the bigram *of course*. So, measures of collocational strength, they quantify the attraction or the predictive power that one word has for the next. And then measures for *n*-grams as I mentioned.

Now about this approach toward validation, why should we validate more, why is it so important especially for corpus linguistics to look outside the narrow confines of it and look at what cognitive linguists are doing. Well, for dispersion or something like that, in corpus linguistics we have now about 20 different measures that quantify that in totally different ways. For collocational measures, there are about 30 measures that are really well established. If you take a recent overview article by Pecina, he actually summarizes 80 different measures. So we have all these. We have different ways to generate *n*-grams like 4-grams, 5-grams, and so on but hardly ever do. People in corpus linguistics now try to find out which one of these are best, which ones correspond to something or predicts something best outside of our own theory. In fact, it has gotten to the really bad state at this point that there are corpus linguists who say "meh, doesn't really matter; just play around with these different ways and then take the one whose results you like best." So why don't we just flush *any* scientific rigor down the toilet? So, obviously we will need some way to validate this type of stuff. It's really useful because different measures yield different types of results. For instance, for collocational attraction, a very nice study by Manfred Krug shows that for grammaticalization processes, what he calls *string frequency*, which is just how often do two things co-occur together, plays one of the most important roles. So he says: ""Well, if you want to look at the coming into existence of something like *sposta*, like 'I am supposed to' or something, you just look at how often does 'supposed to' occur in the corpus." The higher that is, the more likely maybe the chance of grammaticalization. Now in some other studies, some colleagues and I looked at association measures and we found that according to some results I will discuss in a later talk, that this particular statistical measure works best. It doesn't matter right now how this is computed, it's actually a little bit complicated, but the point is just to show, ok for something else, the completely different measure yields a better result. A former M.A. student of mine, Daniel Wiechmann, who looks at the correlation of corpus data and experimental data and found that yet another measure is better. And Dagmar Divjak and colleagues found another measure. So for different types of things we have different measures that yield the best results and wouldn't it be useful if we don't just say "we don't care what happens outside the discourse" but look at the other fields and say "ok what are the things that make us decide between these?" So polemically speaking, should we really just be doing what a lot of corpus linguists are doing and using measures like mutual information or *t* just because these are easy to get, because these are easily implemented and stuff like that? I mean, don't we care that there are psycholinguistic results out there that hopefully affect, first, the choice of statistical measure but then also the results of our interpretation. Obviously, I think we should do that, and that again would mean that, basically, we can't view corpus linguistics as an isolated theory in and of itself. We have to consider a method and then correlate its results to what happens if you apply these methods within other theories.

Now where to turn to and what to relate that to? So like I said, corpus linguistics should be looking at other fields and other results. And why? Well first, it will help corpus linguists themselves and it will help them in terms of the visibility in the field of linguistics and will help them particularly with those subdisciplines of linguistics that have come up with very similar notions, very similar types of hypothesis and data, which I will show you a lot in a moment. Second, because external validation would streamline corpus linguistics research. If, as a corpus linguist, you don't know which type of association measure to use, well then it would be great if you could look at a variety of experimental studies and say "my study is most similar to this and that study found whatever this measure works best, so I am going to use that too." I mean, just to increase your own chances of better results. And that of course then would in turn improve corpus linguistics as a method in and of itself. So we need to hook up as corpus linguistics with other theories. But which other theory could that be? Well, in this context of course it will be kind of obvious but still let's make the point of how do we could approach this. Since Teubert's writing have been so influential for a lot of things that I have been mentioning here, why don't we turn to him for help? So he says, "for me, corpus linguistics and cognitive linguistics are two complementary but ultimately irreconcilable paradigms." He says "Corpus linguistics localizes the study of language, in the *Geisteswissenschaften*, or the humanities." And "corpus linguistics looks at language from a social perspective not the psychological aspects of language" we've had that quote before. If we combine all these pieces of advice he offers, then how about we, as corpus linguists, we turn to something that is psycho-linguistically informed or cognitive-linguistically inspired usage-based linguistics that we locate firmly and deliberately in the social and behavioral sciences? And since we are doing that, since we are talking about humanities and *Geisteswissenschaften*, the German word for sort of *mental sciences* if you will, then of course looking at the cognitive systems, which we are interested in cognitive linguistics, that tells much more about the human condition than if we basically disregard anything that has to do with cognition, which means, I think, at some point, going psycholinguistic or psychological or cognitive, whatever you want to call it, I mean, that is kind of necessary because something only enters into the corpus and then becomes material of these discourses that people like Teubert and others would study if, at some point, it has been processed by your mind, if it has been filtered, and if it has been shaped by some mind, and then has been processed in the way that is also determined by the hearer's internal cognitive or mental structure. Plus, if you look at what has happened in corpus and cognitive linguistics over the last, let's say, 20-something years, the overlap is huge. And in this part of the talk, I basically want to show how a lot of things that you talk about in cognitive linguistics and in corpus linguistics are so closely connected that it actually requires some mental artistry to say these two fields are irreconcilable.

So for example, when corpus linguists talk about token frequencies, how frequently does something happen in the corpus? How frequent is that word? How frequent is that construction? How frequent is this word at the beginning of the sentence, at the end of a paragraph, or something like that? So whenever corpus linguists talk about that, then basically theoretical or cognitive linguists usually become interested in some way, because, all other things being equal, high token frequencies in general correlate with degrees of entrenchment, which Schmid and of course Langacker and other people have been talking about, it correlates with grammaticalization phenomena, phonetic reduction, things like that, and John Bybee, Sandy Thompson, and a lot of other people have talked about this. It also correlates with resistance to morphosyntactic language change, like highly frequent, irregular verbs resist regularization because they are so frequent that they are so well entrenched you don't need to change them in order to facilitate processing. Psycholinguistics or psycholinguists become interested because token frequencies correlate with ease and earliness of acquisition: If something's frequent in a corpus in general then is acquired more early, it correlates with a lot of different types of experimental results like lexical decision tasks, so how quick are you to react to a word on the screen? And you are much quicker to react to a more frequent word than to a less frequent word. From a very simple corpus linguistic finding, you can immediately make psycholinguistic prediction. Now when corpus linguists talk about type frequencies, so how many different words are there in the corpus? How many different constructions are there? How many different words show up in one construction? Then again, cognitive linguists become interested, because type frequencies of this sort correlate with morphological productivity or language change. Things change in language if something becomes schematized because a lot of different things might enter into it. Psycholinguists become interested as well, because type frequencies are also correlated with ease, age of acquisition of constructions, for example. So, if a construction needs some versatility because before it is recognized as a construction, namely that the kid can see, "oh actually there's a lot of verbs that can go into this slot so there is a more general pattern here."

When corpus linguists talk about dispersion like I mentioned before, then again psycholinguists become interested because dispersion has implications for psycholinguistic experiments. So in the study that I will talk about and in a later talk I could show that if you want to predict the reaction times, many psycholinguistic experiments these days use frequencies, but dispersion might actually be more highly correlated with reaction times. So you are not only faster to react to a word if that word is more frequent in general but also if it is more widely used. So the frequency is the same but you have a higher chance of encountering it every day, given the same frequency, you are still faster to react to it. Dispersion also has implications for learning and acquisition. So for instance, Rita Simpson and Nick Ellis have shown that, if you look at applied linguistic data from corpora then academic formulae, things like *on the one hand* and *on the other hand*, things like that, those are better acquired if they are more widely dispersed.

Syntax and lexis, the correlation between words or verbs and constructions. Again, Halliday said "I have always seen lexicogrammar as a unified phenomenon, a single level of wording, of which lexis is the most delicate resolution." Theoretical cognitive linguists completely agree with this type of corpus linguistics statement. So, for example, many psycholinguists have long assumed that words and syntactic patterns, you can consider them to be represented in a mental network and whenever you see something, you're basically activating nodes for words, nodes for constructions, in some ways that hopefully fits the semantics or pragmatic meaning.

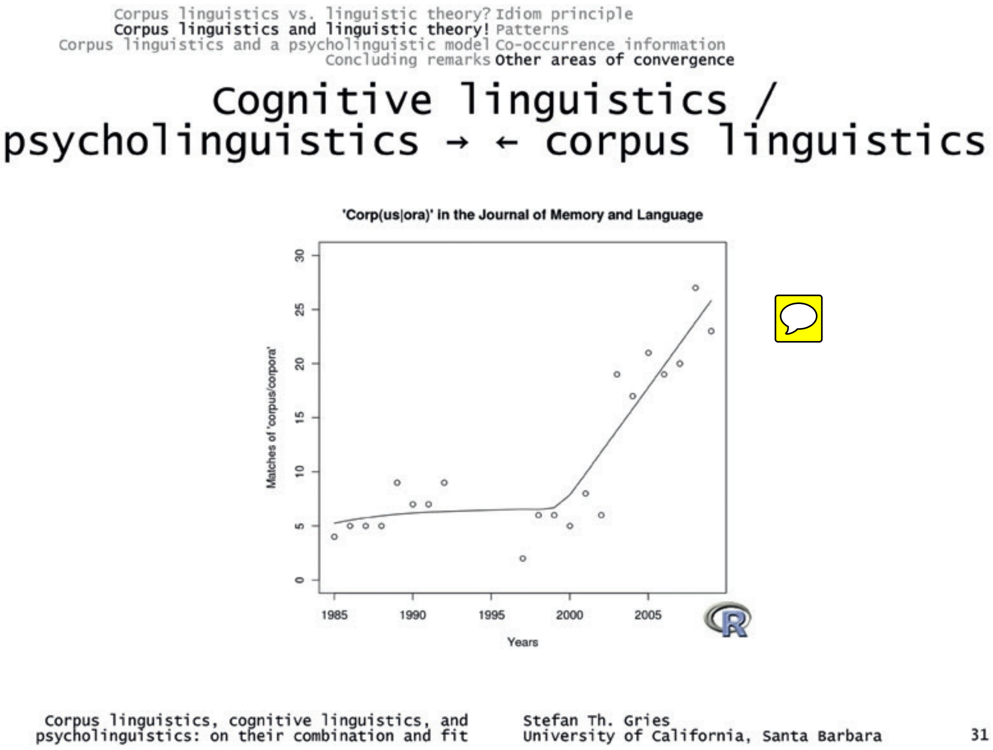
Corpus linguists especially of the British school have been very interested in what is called the idiom principle, which is a principle that states a lot of larger units of language are basically pre-constructed and memorized. So the idea would be in a way, if you say *on the other hand*, you don't construct that word by word. Like, you don't plan to say *on* and *the* and then *other* and then *hand*—you call up *on the other hand* as one word. If you've ever read Langacker, that should sound extremely familiar because, basically, the idiom principle in corpus linguistics says there is a large number of pre-processed, sort of entrenched units, namely in Langacker's term, "a structure that a speaker has mastered so thoroughly that they don't have to focus on their individual parts for arrangement." It is already active as such. And I am going to skip this one in the interest of time for now but the rule-list fallacy would be a similar case where the idiom principle basically makes, I mean, not predictions, but it captures all sorts of observations that are very compatible with this.

Now corpus linguists talk about words and patterns and that should be really interesting for cognitive linguistics because what Hunston and Francis call patterns in corpus linguistics are pretty much exactly what Adele Goldberg talks about in terms of constructions. So this is the definition of patterns by Susan and Jill Francis. "The patterns of a word can be defined as all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it." If that is not the definition of construction, what is? Here is Adele's most recent definition of construction. Note, she actually uses *pattern* here. "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency." So pretty much the exact same thing used in two different theories arrived at independently but meaning completely the same thing.

Co-occurrence information, so corpus linguists talk about concordances, collocations, *n*-grams and colligations. So concordances would be displays of a word how it is used in context. Collocations are co-occurrence of words like *of course, according to*, the examples that I gave, the same with *n*-grams. And collocations are co-occurrences of words on the one hand and constructions or patterns on the other hand. So the fact that *give* likes to occur in the ditransitive, that *go* likes to occur in the intransitive motion construction, and so on. So when corpus linguists talk about all this, then again psycholinguists become very interested in that because such co-occurrence information has been shown to, for instance, help children in language acquisition on the level of phonotactic patterns. So children use co-occurrence information like that to figure out where a word ends, where does the next word begin in an acoustic string. It helps children discover word classes on the basis of words that preceed and follow other words. That's how children seem to begin to abstract away parts of speech. It can predict reading times. So co-occurrence information out of a corpus again helps you make very precise psycholinguistic predictions. Subjects are able to recognize particular *n*-grams faster, even if you control for the frequencies of parts of the *n*-grams. And in general, language production and comprehension has been shown to be extremely item-specific. So, the degrees to which we speed up and slow down as we process or produce language is very context-bound, in a way that is very strongly correlated with the type of frequency effects we find in corpus data.

Now what about one of the other desiderata that Teubert had for corpus linguistics? So remember that he once said corpus linguistics has a social perspective but not a cognitive perspective. Again if you look at recent developments in cognitive linguistics, one of the I think right now strongly growing fields is what you might want to call, what people have been calling *cognitive sociolinguistics*. So people like Dirk Geeraerts, Gitte Kristiansen and others and also more typologically-oriented people like Bill Croft have been arguing for a cognitive sociolinguistics as a field, basically. As you can see, I mean, they are having first conferences, first volumes, special issues; there's just come out a special issue in the *Journal of Pragmatics* and stuff like that. So it's not like cognitive linguists are blind to the social perspective of corpus data as has been implied. In fact, Adele again herself wrote at one point "the function pole in the definition of a construction [so the part that contains semantics in the conceptual constituents of construction] indeed allows for the incorporations of factors pertain to social situation, such as register". So even in the earlier corpus linguistic work that kind of predates the foundation of a cognitive sociolinguistics, the type of social perspectives that people like Teubert want is actually perfectly present in cognitive linguistics. And corpus data in general have become more and more frequently used in cognitive linguistics these days especially if you look at ICLC volumes and the recent volumes in the Cognitive Linguistic Studies series. In psycholinguistics the same is true. So here is a small graph to show this. [Pointing at the slide]

So if you look at the frequency of the word *corpus* or *corpora* in the *Journal of Memory of Language* over the last twenty-something years, then you can see that, for quite some time, the rate was relatively low and has been relatively stable. But at some point of time, it began to increase drastically. So at



Corpus linguistics vs. linguistic theory? Idiom principle
Corpus linguistics and linguistic theory! Patterns
Corpus linguistics and a psycholinguistic model Co-occurrence information
Concluding remarks Other areas of convergence

Cognitive linguistics /
psycholinguistics → ← corpus linguistics

Corpus linguistics, cognitive linguistics, and psycholinguistics: on their combination and fit
Stefan Th. Gries
University of California, Santa Barbara
31

that point one could say, at least based on this little toy data set, that psycholinguists have taken even more notice of what happens in corpus data.

To sum up, corpus linguists talk a lot about stuff that is immediately relevant to cognitive linguistics. And of course I hope to persuade you that that is the with the remaining of my talks to been given here. The sad thing for corpus linguistics in a way, however, is that it is linguists that don't self-identify as corpus linguists so much that apply corpus linguistics methods and that demonstrate that these methods and their results are relevant to things outside of a particular text or a particular type of discourse, [...]. So a lot of times, it is people who are not corpus linguists who validate the suggestions that corpus linguists have made because it is people outside of that field who actually make the external connections and see to what degree particular corpus linguistic notions, hypotheses, and data and so on have something to contribute. So we as corpus linguists can benefit a lot by looking at what happens in these oh-so-different irreconcilably different disciplines because these disciplines have developed tools and they have developed methods or theories or frameworks, if you will, that allow corpus linguists to move away from the often criticized purely descriptive number-crunching type of what they do towards something that is interesting and the explanation of linguistic phenomena with regard to something outside of corpus and embedding things into a larger context. By a larger context, I mean theoretical context.

As I will try to argue here in a moment, is that the type of psycholinguistic approach that corpus linguists and a lot of cognitive linguists that don't think about frequencies that much would want to be interested in is an exemplar-based approach. Now what do I see here as an exemplar-based approach? Well, basically the way I want to explain this starts out again from first language acquisition, basically starting out from the observation that children, I mean infants, toddlers, are extremely good at keeping track of the distribution of characteristics of the ambient language. So children can listen to even meaningless sequences of sounds over the course of just a few minutes and find distributional patterns in there that help them to see "where do words begin, where do words end?" The question now is, how does that happen, and how is it represented? And again the answer will be something like exemplar representations. Now what are the main assumptions of this model? In a way ironically enough, one of the best quotes defining this is not from a cognitive linguist but from an 'irreconcilably different' corpus linguist, Michael Halliday, who said "each instance [so each perception or production of a particular word or a construction or something in the context] redefines the [linguistic or cognitive] system, however infinitesimally, [so however small the impact is], maintaining its present state [so not changing it at all] or shifting its probabilities in one direction or the other." So if you hear me using a particular word in a particular way that you haven't heard before, then that might or might not have a little bit of an impact on your linguistic system, on your knowledge of English, or whatever language someone else is speaking to you and it changes the system a little bit. And if you hear it again, in a relatively short period of time, you might just learn that; if you never hear that again, you might forget it again. But even hearing something for just a single time might already have a long-lasting impact on your linguistic system.

In another way, this was put very nicely, this time from a cognitive linguist, I think, he would be happy to be called that, Nick Ellis, in his very nice overview paper on frequency effects: "it is usual that each learning event updates a statistical representation of a category independently of other learning events." So, again, as soon as you hear in something, the use of a particular word in a particular context, your linguistic system gets updated because you say "oh that's possible too". So you assign a probability to it that is suddenly greater than zero. Before that, you might have thought, "ok, that can't happen." But then you hear, in this case, a native speaker use it, and you think "ok actually it *can* happen" So I can say that too at some later point," and your system has been changed a little bit. So that means speakers and listeners, they remember aspects of tokens and place them sort of into a multidimensional space or network. The *aspects* here are important. So the idea is not "you remember everything you've ever heard and all particular aspects in details of it" but over time of course we generalize, we don't pay close attention to everything. But some things may get lost very quickly and in fact it was Nick Ellis who argued against that was actually necessary for generalization. So, for example, labels such as those of phonemes are associated with a distribution of memory traces in a parametric space, and I will show you a graph in a moment that highlights what I mean, and in this case a cognitive representation of the parametric phonetic space. That's a quote from Janet Pierrehumbert, who for instance showed this for phonemes with a well and often-cited graph like this, where the idea is that on the *x*-axis we have the frequency of F1 [pointing at the graph and describing it], the formant frequency of a particular sound. On the *y*-axis, we have the frequency of F2, and then each of these little symbols here represents a particular sound that you heard. So this sound you heard have an F1 frequency of, whatever, like 350 or something and an F2 frequency of, let's say, 800. And then you heard this sound which also had a F2 frequency of about 800 but slightly higher an F1 frequency and so on. Then over time what happens is that basically we build these clouds of sounds that were relatively similar to each other along these dimensions. Then, over time, we group them, as circling them as belonging together, so you can see here—it's not the best resolution—but you can see sort of different symbols have been used for different phonemes. So these are obviously vowel phonemes, so the "æ" sound. So this would be the core area of the "æ" sound. But crucially note, for instance, here is an "x", so this an "æ" sound that is actually in phonetic space that is closed to the "ʌ". But in that context, we nevertheless interpreted it as "æ", I mean although it looks like an "ʌ" sound in terms of its phonetic characteristics, because of the contexts, we still make it part of the cloud of the "æ" sound. So whenever you hear a new instance of a sound, you basically plot a new point into that system and everything changes a little bit. So if the next "æ" goes up here, then that might mean you make the circle a little bit wider and you update your statistical system; your statistical knowledge in this case for that phonetic space a little bit.

Here is a different example in terms of perception, if we have F2, this time on the *x*-axis. And then we have the strength with which you have particular F2 values memorized, then you might have one type of volume represented here by those straight lines. This is a less frequent vowel, let's say. And then you have another vowel represented by this dash lines, that's more frequent.

And then there might be an incoming sound which is located at the "x". And then your task as the hearer basically is to see, well, which of these two is that? Is that this one? I mean it's a little between the two, right, so you have to make a choice. So, is it more like this one, or is it more like that one? And depending on what you do, you will either draw a little dashed line here or a little straight line over here and change your overall knowledge representation of that system. With that perspective, I think it's kind of clear how corpus data basically provide a lot of relevant information. So these co-occurrence spaces involve all sorts of possible dimensions. So phonetic and phonological information of the type that I have just shown, but then also prosodic, morphemic, lexical, syntactic co-occurrence, and all sorts of extra-linguistic aspects as well, utterance context, sociolinguistic speaker factors, register, genre, and mode and so on. For example, if you are a native speaker of a language, but you hear a non-native speaker of that language talk to you, then you very quickly adjust your system to how that person speaks your language, right? So that you know that person might not get the vowels right in exactly the same way that native speakers would but often after just a few seconds, you already adapt to it and you already process that different type of input. And why is that? Because in this case of utterance context and characteristics of the speaker. So in a way we are able to tweak the system or listen to input to the system in ways that move these dimensions around.

But like I said, it's not like we represent everything or remember everything. Some aspects of individual tokens of characteristics might not be accessible or might have never entered into the system. Ewa Dąbrowska at one point said "How are learners of a language then able to isolate typical contexts for a particular word?" and then seemingly paradoxically, but I think she is right, she blames on the fallibility of human memory: the fact that we normally don't remember things we encounter only once or twice in a way it helps. So some things we never store in long term memory, and some things may just decay and we may generalize or abstract over them. As Nick Ellis said, "abstraction is an automatic consequence of aggregate activation of high-frequency exemplars," so if you hear particular sounds or particular words a lot of times and you don't remember every single characteristic or every single instance anymore, you generalize, which means you form a larger category with regression toward central tendencies. So you build a statistical average of all the different examples that you hear and that may well become the prototype at some later stage. Unfortunately at this point, I cannot use my own computer; I had a 3-D simulation to show you how that might work but maybe we can get that set up at some other time.

Why will models like that be great? Apart from the fact they obviously have a strong correlation with corpus linguistic approaches? There are some theoretical advantages. One is that if you adopt this approach, you can talk and explain a lot of things about our first language acquisition without regard to anything like untestable, for example, generative grammar type of parameters. You don't have to assume that the child has innate knowledge that languages are head-first or head-final. You just say the child abstracts that out of all the stuff that he or she hears all the time, constantly updating the linguistic system and that kind of helps. Second, we know the speakers and listeners store immense amounts of probabilistic information anyway. So we've known that from other case: 'it's not like we are making a particularly weird assumption—we are just using knowledge we have anyway. And this type of approach that says we store information in a multidimensional space can explain these types of frequency effects really well. For example, in this type of approach you would say that high frequencies of occurrence of something corresponds to a dense cloud with many different points in close proximity. Remember Pierrehumbert's vowel-based type of graph? Some of the vowels are really frequent so these points clouds in F1 and F2 space, formant space, they are very dense and if something is really dense, then we're better able to recognize it really quickly, same thing for high frequencies of co-occurrence. If the verb *give* shows up in the ditransitive a lot and both of them, on the semantic axis, mean transfer, then there is a very high probability for us to recognize that co-occurrence and we look at the verb and the construction then from that type of semantic angle. Same then for categorization and prototype effects. If you remember the Ewa Dąbrowska and Nick Ellis' quote: over time we hear things a lot of times and we may make generalizations, we abstract away to an average type of conflation of a lot of different characteristics. That might very well be what gives rise to the structure of category as we extend it and maybe modify into a radial category of the Lakovian type of approach. And it gives rise to prototype effects as we recognize "this thing is really close to the middle of the 50,000 exemplars that I have 'saved', so we are fast at recognizing it. More interestingly, especially because this is not often talked about but I think it's really important, this type of approach can also explain how it is the case that even native speakers of a language are very different in how good they are in their language. So again it was Ewa who showed, in cognitive linguistics at least, there are huge individual differences between how well native speakers of a language can process their own native language. She found that professors at her university were better at dealing with multiple embeddings than janitors in her university because one of the two groups obviously has worked with language more directly, more hours of the day, has more interaction with people who have a particular degree of syntactic complexity and so on. And right now there is a student at Stanford who works on something like this too, who basically looks at how, for instance, length differences that help explain order alternations, are different across different individuals. So one hypothesis would be, if a processing difficulty is a function of how large your working memory is, then you should be able to predict how people would express particular states of affairs depending on their working memory capacity. So if you have a multidimensional exemplar space model, then you don't even have to explain this, it's kind of natural that that would happen, because all of us hear different things every day. Of course, all of our systems will be slightly different. So you don't have to jump through extra hoops or postulate why that might be the case or postulate whatever performance factors or something like that. It just falls out naturally from the model you implicitly assume anyway. Yeah let me skip that.

What are the methodological implications of this type of approach? This is something I will be talking a lot about in the coming days because of the statistical theme of many of the talks that I will give. It means that we have to look at the corpus data in a multifactorial way and a lot of time at least in a hypothesis testing way. We will talk about how multifactorial regression models or other types of statistical approaches will show which dimensions of the data we look at are actually relevant and are probably retained or processed by speakers. Here are some examples of studies, several of which will come up later. In studies of alternation phenomena, like the dative alternation, the difference between *he gave him the book, he gave the book to him*", particle placement, *I brought the book back, I brought back the book*", so many of this has been looked at from the cognitive linguistic perspective with this type of statistical tools. We probably will need more what are called mixed-effects models, so regression models that are specific and detailed enough that they allow us to take the abilities and the linguistic systems of different speakers and the effects for different words into consideration. So we will be able to not just generalize over all sorts of different speakers, basically forgetting about the individual variation, but we should be using tools that can distinguish, well for instance, to what degree speakers, maybe different types of working memory capacities, might have an impact. It will require more bottom-up and more multivariate methods to see which dimensions, which kinds of information in the space are more relevant. Again, I can unfortunately not show you the graph that I wanted to use to exemplify this point. And these will be methods that go on that direction.

To wrap up, what I wanted to do in this first talk that basically foreshadows many of the things I will talk about in the subsequent talks this week, I wanted to discuss some of the reasons why theoretical linguists in general, but also corpus linguistics, in particular a particular school of corpus linguistics why they haven't interacted in the way that I think they should have and why I want to see that more. I wanted to talk a little bit about why I think this gap between these fields should be closed at a much faster pace and why corpus linguists especially need to learn a lot from cognitive linguists in the coming years. And then I wanted to convince you especially if you haven't worked with corpus linguistics so much yet, that much of what is happening in the field and what had been happening in the last 15 or so years is extremely compatible with cognitive construction grammar of a Goldbergian type of flavor. Remember all those slides where I showed you that while corpus linguists do that and it has exactly an analog in the cognitive or psycholinguistic approach. And that I think corpus linguistics questions can then be answered in much more revealing ways because we can suddenly look at cognitive mechanisms explaining some things that otherwise will be unexplainable by looking at the corpus in and of itself. For example, wouldn't it be nicer if we can explain particular distributions in corpora, such as "oh, the production of this word is reduced" with reference to cognitive mechanism such as learning, habitualization, and articulatory routines rather than just say "well this is what happens in this discourse?" Or wouldn't it be better to talk about grammaticalization as cognitive linguists do, but many corpus linguists still don't do well enough? So they look at changes in diachronic corpora and it would be nice to be able to say this happens because of automization and that happens because of entrenchment and that happens because of frequency of occurrence?

What I want us to do basically is to first rethink this notion of corpus-driven and corpus-based linguistics. And those of you who sort of look more into corpus linguistics in time, I would really advise that, basically, just forget about it. I mean just take my word, at least at the beginning, for it until you've read more yourself and realize corpus-driven linguistics are really in a way that should be done according to their proponents that doesn't happen, that corpus-based linguistics might be more useful. And we should definitely rethink that whole warfare rhetoric, about how we can't work together, and how we talk about all sorts of different things because I have tried to show you over several slides that that is actually completely untrue. But the main proposal is that corpus linguists should adopt the cognitive perspective, and one that is inspired by usage-based linguistics and exemplar-based approaches. And of course I am not completely the first person to say that. There is one quote I want to bring up here and another favorite quote at another point later in time. This is a really relatively early psycholinguistic paper that looks at the stuff like that. Miller and Charles were looking at antonymy and synonymy and things like that and argued for what they call a *contextualized representation*. They basically have a completely psycholinguistic goal but the way they defined it actually was very corpus-heavy. This is here is the favorite quote actually from a corpus linguist who, funnily enough at the time he wrote that book and developed that theory sort of always thought that psycholinguistics would really be cool to look at it but he never had the time to do it. But it's interesting the way that he defined his work and that is by this and I really think it's worth quoting that verbatim: "The mind has a mental concordance of every word that has encountered, a concordance that has often richly glossed or annotated for social, physical, discoursal, generic, and interpersonal context." This is a progressive corpus linguist's view on how cognitive corpus linguistics and psycholinguistics basically all interact. I think it's exact the type of perspective that we should be taking, and that I hope I will advertise here in the talks to come. So some people are on that track but the major breakthrough, I mean, the greater recognition, this is what we need, has unfortunately not yet happened. Thank you.

LECTURE 2

# Quantitative Approaches to Similarity in Cognitive Linguistics 1: The Phonology of Blends

The next three lectures of mine basically will be on different types of topics and they were chosen, basically, with an eye to exemplify a few ways in which corpus data and statistical methods can be used to study things that a lot of times people either said they are really difficult to study at all or you can study this type of stuff but you can't do it well with corpora or you can't do it well with statistics and things like that. So basically what I want to do in the lecture today and the two tomorrow is to highlight a few ways in which more or less advanced type of statistical approaches on corpus data can help to shed light on a variety of phenomena otherwise difficult to study. The first one today, basically as you can see, will be concerned with blends and some aspects of their phonology and then the other two, to just give you an idea already, will be on idioms and corpus-based cognitive semantics tomorrow.

Blending is a relatively perplexing word formation processes, and that has a variety of reasons. One is that it is not as rule-governed as many derivational morphological processes are. It's not like you can characterize what happens there with a few relatively exceptionless rules. Second, it's not as productive as most derivational processes are. So in English most of the time you want to form an adverb or something you add *ly* maybe to an adjective or something like that. In blending it's really very different. It's also much more creative so you can do a lot of different things with blendings, as you will see, whereas derivational process, given their rule-based nature, are much more rigid. And also they usually involve a conscious formation. Apart from speech-error blends, which I will talk about only occasionally, most of the time if you create a new blend, that's a conscious effort. You decide on how to put together two words and I will show you some examples in a moment. Whenever you as a linguist want to try to analyze how people form a blend you can't just fall back on unconscious linguistic processes but there's also a lot of conscious processing going in and that of course might differ a lot across individuals. So if you decide how to blend the words *breakfast* and *lunch* into *brunch*, then there is a variety of ways you can do it and there is not a single obvious answer. Now the fact whoever did that at some point made the word *brunch* doesn't say that that's theoretically the optimal way to blend these two words and it's really too hard to figure out why exactly they did this in this particular way.

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_003

There are some clues, but I mean, one thing for instance that I want to mention as an example already, is that in some sense, *brunch* is not a really good blend because if you have never heard the word and you hear it in the context, then the context might well involve food so you might guess *brunch*, that word *brunch*, you don't mean something with *lunch*. But it's difficult or more difficult at least to figure out what the first word is because there is a large number of words that begin with *br*, so figuring out from the word *brunch* what that word might be, that first word, is difficult. In one paper title I am sort of jokingly arguing that *breakfunch* would actually much better as a blend because it still ends with in *unch* so people can recover *lunch* really quickly but then it also begins with *breakf* so that's a dead giveaway for *breakfast*. So in a way, in terms of recognizability, that would be a much better blend. Anyway, the point being, like I said, that people who coin these formations, they often give it a lot of conscious effort and thought to figure out what are the most memorable or funny or whatever type of information will be and that of course makes some difficult to analyze. Especially what we sometimes find is that they also result in violations of morphological rules. For instance, most blends consist of parts of words like *breakfast* or *lunch*, that sort of respect phonological entities. You don't often find blends that split up the onset of a word like a cluster. So whatever, *bunch*, for instance: no one, I guess, no one would form the blend of *breakfast* and *lunch* by just taking the *b* of *breakfast* because that will split up the *br*, the consonant cluster at the beginning. That's really rare but actually there are cases where that happens. If you have a large enough selection, then you will find variety of really weird things that people have done because I guess they thought that's funny or something, so that makes it hard to explain and then you need to basically integrate a lot of different types of information, so some blends play with the fact that, the way that they are spelt, the way that they are pronounced ... Really you only see it's a blend if you see it in spelling. There are really some blends that I can't think of an example right now where if you hear a word, it sounds like a normal word, but actually, but when you see it spelt, you see this is differently spelt from that word so it is actually a blend. So you have to integrate phonology or pronunciation on the one hand and spelling or orthography on the other hand, which again makes it more difficult because of the high degree of arbitrary conventionalization in spelling.

There are some similarities of blending to other processes, for instances, compounding because, obviously, compounding usually fuses two or more words, as do blends and then maybe also, what some people said is actually the same, and I will show you that's not the case, namely complex clippings. A complex clipping would be a case where you merge the beginnings of two words. So one example that many people know would be *scifi* for *science* and *fiction*. I mean the *sci* is from *science* and *fi* is a slightly altered part, a version of *fiction*. So you have the beginnings of the two words where in *brunch* you have the beginning of one word and the end of another. Or in *motel* you have *motor* and *hotel*, the beginning of one word and the other. There are some formal similarities to these other processes but, as we will see, also some differences. Blending has an unplanned counterpart, namely when people commit speech errors. Sometimes, not often, but sometimes, the commit them in such a way that instead of saying one word, they say two words that pretty much mean the same type of thing and they merge them accidently. So one famous example is *absotively* as a blend of *absolutely* and *positively*, which in certain contexts can mean the same thing. It's a really tricky process that involves interactions of a lot of different things on a lot of different levels and because of that, for a number of years, there is a really pessimistic stance in morphological research on that and I have two representative quotes here. So in a very well-known Cambridge textbook Laurie Bauer wrote: "In blending, the blender is apparently free to take as much or as little from either base as is felt to be necessary or desirable [...] Exactly what the restrictions are, however, beyond pronounceability and spellability [if that applies] is far from clear." Basically saying, "well, we have no clue." That's the scientific way of saying, "we don't know." Similar pessimistic quote, in a way, "we find no discernible relationship between phonology [...] and a viable blend. [...] This fact helps to make blends one of the most unpredictable categories of word-formation." And in a way, it's true, I mean, they involve a huge degree of complexity because of all these different potentially conflicting factors and it's really hard to come up with solid and statistically robust generalizations.

However, there is a different side to that and that is, just because blends don't exhibit many of the categorical features doesn't mean they are unpredictable. It's just that we basically only have to put ourselves into a position where we recognize that, actually, in linguistics, a lot of things are not completely predictable and categorical. Pretty much everything is probabilistic, everything is multi-factorial, determined by a ton of different features at the same time, so actually that's not a big surprise that blends will do the same thing. It still makes it difficult but not a particularly tricky exception in a way. And so we need a probabilistic approach to deal with this. And of course a cognitive linguistic approach or psycholinguistically-based approach would be something like that. For that, of course, it also means we need a large samples of blends. Many studies until relatively recently ago the sample sizes were really quite small. So here are some examples: Pound (1914), in other ways a really nice early study, she has 300 blends, Cannon, a formative article, had 132 and a variety of other cases like this; I will mention some more later. So obviously it's very difficult to come up with good generalizations on the bases of such small samples. And secondly, we need statistical methods then, obviously, that can handle uncertainties, that can quantify patterns in a way that goes beyond what you can just do when you intuit on what might be going on in a particular data set.

Here is the official definition of blend I will assume, so it's a fusion of two or more words, and I will focus here only on cases where we have two-word blends, so there are a few cases where people joined three words but most of the time it's two only, like more than 95% I would think. Two or more words where the part of source word one, so the first source word, is combined with a part of source word two where at least one source word is shortened, so loses some material, and/or the fusion made involves overlap of the two. If you have a blend like *foolosopher* then that's a blend of *fool* and *philosopher* and they share the *l*. So *fool* has actually not been shortened, I mean *fool* is still in *foolosopher*, but the *l* overlaps with the *l* of *philosopher*. That would be according to this definition, that will still be covered as being a blend, or if you have *motor* and *hotel* in *motel*, then in spelling they share the *ot* obviously.

Now what I want to do here then is talking about several different case studies, like I said before, highlight the ways in which statistical analysis based on my collection of blends and also additional corpus data that I will talk about later helps us see that there are actually a few patterns in blending that make it less unpredictable than has been assumed. I will, just for the sake of heuristics here, I will divide the talk into three different parts that have to do with the three temporal, or reflect the three temporal, stages of blending. So, the selection of two source words if you want to create a blend, what do you have to do? First of all, you will have to select two words that you want to blend. Then you have to decide in which order. If you want to join *breakfast* and *lunch*, I mean, temporally speaking, it's kind of natural that *breakfast* will come first because in the normal day it comes first—theoretically, it could be the other way round. Then, once you have decided on that, then you can decide on how to blend them. One example that you will see later in the slides, for instance, will be from pseudo-experimental setting, if you take two American car brands like *Chevrolet* and *Cadillac* and let's say, the companies were to merge, then you want to reflect this in the blend then how do you merge them? You could say *Chevrolac*, that would be one blend so I mean the two words are given it's got to be the two company names. Then, you can decide I put *Chevrolet* first, so it is going to be *Chevrolac* or something like that but you can also say, I put *Cadillac* first so you make it *Cadillet*. So you have to make that decision what come first and depending on what comes first, then you have additional decisions on how to order them. So for instance, once you decide *Chevrolet* comes first, you could say *Chevrolac* but you could also say *Chedillet*, like cut it off a syllable earlier something. These are kinds of things basically you have to decide. I am not saying there is a speaker who creates the blend necessarily goes through these stages as systematically as I am discussing them here but for for organizational purposes, I think it's a good scheme to split these phases up. And for these different case studies we'll have to look at many different phonological analysis like I've already mentioned you talk about graphemes and phonemes because some of the blends only show up or are only recognizable in one of these shapes. Also you will see the graphemes, phonemes and *n*-grams play a role. We need to talk about syllables and their constituents because not all blends are created equally in that regard. And then of course obviously we have to talk about words and some factors that determine these things such as the word lengths, the word frequencies and their semantics at least to some extent.

In terms of methodological considerations, I think one thing that is really important in this type of study is that you need to compare whatever you find for intentional blends to other formations to see how they differ in terms of selection, ordering, and blending. So I don't think it makes a lot of sense to talk about blends in isolation, it's always good to talk about them either with regard to, let's say, complex clippings or compounds or in terms of their closest siblings in away, namely error blends so what happens if people do this type of stuff, but they all do it unintentionally?

Second, a lot of times as you will say we need to talk about testing whatever we find for blends against random baselines because just because you find a particular percentage of blends exhibiting something doesn't mean that that is already noteworthy because maybe every word exhibits that to some degree. We need to compare whatever you find for blends against what you might find on the basis of a random distribution. As you will see, we will need successive fine-tuning of methods, so I will give you at least one example where one way that I suggested to do stuff actually is wrong and so today I will show the wrong way, and I also think of a better way and this is probably still not the final word of how this should be done, although I think it's much better than before. And then we need hopefully larger data bases. Right now the database I have has about this size here. It's about 2330-something formations that includes intentional blends like *brunch, motel, foolosopher* and stuff like that, it includes error blends like the *absotively* example that I mentioned, and it includes complex clipping like *scifi* or *sysadmin* for *system administrator*, things like that. These blends were all annotated for a huge number of features so that's where these data points come from. All of them were manually annotated, for instance, for the frequencies of the words, for the semantic of the words, for how many letters, how many phonemes, how many syllables from one element went into the blend, did not go into the blend. Where the split-up happened, between the syllables, between the onset and the rhyme, between the body and the coda. This brings us back to one of the questions this morning, do you really have to manually annotate all the stuff? Yes, you do!

This is kind of a schematic overview of where we are, where we're going to go. So again we kind of have the three heuristic time periods in the creation process: selection of words, ordering of source words, blending of source words. Here is a variety of things I looked at in the past few years. We are basically going to have to look at the blue, the stuff that is highlighted in blue. One thing that already I want to bring up here is that there are two main forces, I think, that are useful to distinguish when it comes to how blends are created and those are the ones I listed here on the left hand side, namely similarity and recognizability. Similarity plays a role in terms of that we will find that words that make it into blends are often more similar to each other than what you expect by chance. Plus there's the additional question of how similar is the blends to each of the source words, which in turn, and that's why there's these double arrows here, increases its recognizability. But as you will see later, those can also conflict with each other. You can make these two source words super easy to recognize but then the resulting blend is actually a really bad one, because it's not similar to either of the source words anymore, and I know it sounds like a contradiction, but you will see it's not.

This is what we are going to start with, the selection of the source words that are going to the blend. As I have said, there's a variety of studies that have shown that if people choose to coin blends, they often choose words that are similar to each other. And part of that of course has to do, a lot of times you don't want to just merge two words to create a new word. If you want to do that, you can just do a compound, the reason then to do it with a blend is, because the blend is funny or witty or something and maybe more memorable. The same thing actually seems to hold for speech error blends, and on a variety of levels. This is again, like I said, where the problems start: You have to keep so many things in mind at the same time. So for instance, we know that, if you look at speech error blends where people mix up two words, those are phonologically similar to each other. They sound the same, usually they have the same part of speech. So, syntactic characteristics here basically is a short hand or a different way of saying there are the same part of speech: People don't usually accidentally blend a noun and an adjective. They blend two nouns that mean the same thing into a new noun. They blend two adjectives, but they don't blend an adjective and a noun. And they are often similar in terms of semantic characteristics, so a lot of times, the two words that are blended accidently are near-synonyms or complete synonyms even. We also find that for intentional blends, though to a much lesser degree. But then there are many different ways in which words can be similar to each other. Here are some examples and we are going to look at several of these. So one would be length, another one would be frequency, then phonemic and graphemic material; this would just be, for instance, do they have the same number of letters or phonemes, but then this would look at how many letters or phonemes do they actually share and maybe in what positions. Then stress patterns, so just is the word 'stressed-unstressed-unstressed', or are both [words stressed like] that. And then finally, a sort of very coarse-grained, and at this point a very preliminary, study in terms of semantic features. We will also see that words can have different places where they are similar to each other and, finally, that makes it even more difficult again, there are more different ways to measure these similarities, some are relatively straightforward, I mean if you look at length, then you can just say this word is seven phonemes long and this word is five phonemes long so there's a two-phoneme difference, but with other features, that is actually not that straightforward.

Let's look at something very simple. First, namely, the example I just gave, length of source words. For each of the formations in my database, I determined their lengths. And as you can see this is already tricky, because you can choose at least three different units. You can take the most coarse-grained one—that would be syllables—but you then can also take something that is much more fine-grained, which would be phonemes and you can take graphemes/letters and of course you all know in English the phoneme-grapheme mapping is not exactly ideal in terms of one-to-one mappings. So these results might often seem to be on a similar level of granularity but actually they can differ considerably sometimes. And then for each of these forms in my database I also determined what type of formation it is. For this little case study here, I will just distinguish these three different forms: One could be an authentic error blend. To just mention where those come from: In the 1970s, there was a lot of psycholinguistic research on speech errors. I guess a lot of students and a lot of linguistics or psycholinguistics professors were perhaps running around basically noting everything down when people made a speech error, noting down the circumstances of production, the intended target word, and then what people are actually saying and so on. So that's where the authentic error blends come from.

The induced error blends; there was a short period of time where people did . . . basically put people in labs and try to get them to produce error blends accidently. One way to do this is to put people in a situation where they would have to . . . It's like tip-of-the-tongue states. So you put people in the situation where you make it very likely that they produce a particular word but maybe in the context where you hear another word before and the two are semantically similar, and you basically hope that a lot of times people will commit an error, which is of course extremely labor-intensive and sometimes not really productive because in general we don't perform many of these errors all the time, so you have to run a lot of subjects and put them in a lot of different situations to get decent error blends. But to some extent, that worked so that's where they come from. And intentional error blends are taken from a large number of websites, pretty much everything that I've read on blends, and then I had some people, whenever they saw a blend and sent me an email and said, these are source words and then I added it to the database. Like I said, I have some other formations, and they are like complex clippings but in this part I will not talk about those too much.

When I did this, I compared the length of the source words to each other for each of the blend types and one way to represent this is like this. And maybe I should work through this a little bit. This is a so-called box plot. We have three different panels. As you can see, we have one for the authentic error blends, one for the induced error blends, and one for the intentional error blends. For each of these little panels, we have a statistical results for the first source word and one for the second source word. Then we have, in this case, the number of syllables on the *y*-axis. So one, two, three, four up onto seven syllables. Then in these box plots, this heavy line here is the median, so one particular type of average, the exact statistical nature of which is not relevant right now. If these—these are called *notches*, these little things that lead to the heavy line—if those do not overlap like they do not do here, then the medians are mostly likely significant different. So we can see, for instance, what this graph shows then, as big and complex as it is, the main thing that it shows actually is that, in authentic error blends, people really commit an error in real life, no experiment, nothing, then usually the source words have the same length. That seems to be something that makes people conducive to producing errors. In induced error blends, that is true too. However, this has to be taken with a caveat because here it was researchers that picked the source words because they want people to make error blends. So we have to take this with a huge grain of salt. But then the crucial result here is that apparently for actual intentional formations, there is a difference between the two: The second source word is longer. In that case, *brunch* would actually not be a good example because in *brunch* the first word is *breakfast*, and it has two syllables, and the second word is *lunch*, which has one. So in that case, it's the other way round whereas in most cases on the whole, you find the first source word is two syllables long and the second one is three [words long]. And that's a hugely significant difference. There is no way this happens by chance, there is definitely a pattern there.

Just to show this is not an accidental finding, in this case we find that for the other operationalizations of length, too. So here we have the same plot: authentic blends, induced blends, intentional blends. But this time, length is counted in the number of phonemes. And again we find that if people commit errors, then the words are usually exact the same length so there has to be something: if the two words are competing to be produced have the same length, then they are more likely to end up being blended. But when the blend is intentional, then again the first word is usually shorter than the second one. So the first word has a medium length of five phonemes, which was two syllables and the second one has a medium length of like seven phonemes or seven and half phonemes, which was three syllables. Same thing, a rare case where everything coincides I guess, for graphemic length, so now looking at letters and characters: intentional blends, same length, although that actually shouldn't even matter because usually when people speak you don't think about how stuff is written, right? But when it comes to forming intentional blends, the second source word is longer. So a really robust trend that we see in all three cases. The source words of error blends are short with the same length, induced error blends are longer but not different from each other, but again that is more a reflection of what people chose to put in their labs, and then intentional blends are differently long: The first source word is short, the second one is significantly longer. What does that show? It shows the source words of error blend is different from the source words of intentional blends. That may sound like "well, duh", but there are people who have argued that intentional blends are basically a product of the same process than error blends and it seems at least that the selection process, so to speak, is different. And it also means that finding from induced error blends actually had to be really taken with a grain of salt because they may tell you more about what researchers thought is going on than what actually people are really doing with blends.

One thing that is worth pointing out here is these are not pairwise comparisons yet. What I mean by this is that these plots here [pointing at the ppt], that this comparison here, is just a sort of all first words, all second words—it's not the comparison of the first word of a blend to the second word of the same blend. We will get to this a little bit later. There's *breakfast* in here, there's *motel* in here, but this is not compared to the *lunch* and to the *hotel* here. It is not a pairwise comparison within the same blend—it is all at the same time. We will get to the finer level of resolution at a moment.

Now what about frequency? Length and frequency are of course highly correlated. So it may seem as we can expect what we are going to find there and in fact we can. And in the handbook those are on different pages so that the graph would not overplot text. So again we find in authentic error blends, the two words are equally frequent. And the same thing in induced error blends, but if we look at intentional blends then what happens is that the first word is more frequent than the second. And that makes sense because we found that the second word is also longer and longer words are in general less frequent in the language. So it seems that what people are doing is that they put the more frequent thing first, which then happens to be shorter. In a way, from the length result that was predictable.

But now what about similarity, how similar are the two words that you pick with regard to each other. How similar is the word *breakfast* to *lunch*? That is already not a trivial question because there's a lot of different ways in which that can be operationalized. What I want to compare here are error blends, both authentic and induced, to intentional blends but then this is a case where we also need a random baseline, namely I want to compare attested blends, so blends that actually do exist, to randomly-chosen words. Now why is that the case? Because ... I mean just think about what you get as a result if you don't do that. Well then you might say I have ... Ok, let's say you got a thousand blends/ Then you compare, for the one thousand blends, you compare how similar is the first source word to the second and you do that with some number. Then, you compute the average for the set of one thousand blends and then you get the number point two, and what does that mean? Actually that means nothing because you don't know how similar words.... I mean, any two words can be similar to each other just by chance. You just pick two words. Chances are they share a phoneme or two. Just saying how similar the words of blends are to each other actually doesn't do anything for the analysis of blends. You have to compare it to how similar are words to each other in general. That's why there will be a baseline where I basically took random words and compared their similarity to each other. Specifically, to mention that already, I don't think it shows on the slides, so I took one thousand words from the CELEX database and then computed the similarity of each of the one thousand words, for each of the one to the other 999. The first word was compared to word two to one thousand. The second word was compared to one and three to one thousand, and so on. So if you do that with one thousand words you end up with half a million comparisons. Obviously I didn't do that manually but with a script, but just saying. The question then becomes how do we measure similarity? One way to do this is this. This is the simplest one with what is called the Dice coefficient. So it's the percentage of any type of shared bigrams, where *bigrams* here refers to two-character sequences. So let me show you an example, there is a blend *chunnel*, which is the blend of *channel* and *tunnel*. And it doesn't take a linguist to see that those two words are similar but it takes a linguist to quantify *how* similar they are. One way to do it would be with the Dice coefficient where the idea would be: you write down every word in terms of these bigrams, *channel* as you can see, I am using letters here and you can see, you can also use phonemes: so it's *ch, ha, an* . . . you get the picture, and then you count how many they share. So in this case, they share the highlighted six out of all eleven so the Dice value here is 0.55. A probably better approach would be what is called *string edit distance*. That works like this. You take the two words and then you check how many changes you have to make to one word to transform it into the other. Basically, here you have to delete the *c* because it doesn't show up in here. Ok so that's one step. Then you have to replace the *h* by the *t* because there is no *h* in there but the *t*, that's the replacement operation. And then you have to replace the *a* by the *u* and then you are done. So it takes three steps to make this into that. That would be the string edit distance.

Now one nasty thing about these two measures is that the Dice value reflects similarity so the higher it is, the more similar these two words are. Now this one reflects *dis*similarity: The higher this is, the more *dis*similar they are. I mean there is just no way around that. Basically you have to bear in mind, this is actually similarity measure, high values means high similarity, this is *dis*-similarity measure, high value means high *dis*similarity. If two words were the same, string edit distance would be zero. You don't have to change it at all; it's done already. So in this case, I will only talk about the phonemic description of source words. But on a relatively large database, 180 error blends, a ridiculously small number of induced error blends, and I think a relatively larger number of actual intentional blends. This is one way of summarizing these data. This is a very useful but, unfortunately at the same time, extremely unintuitive plot. So let me explain what that does. Let me explain it on the basis of this graph; so this is for the string edit distance. Ok, so this kind of plot on the *x*-axis you have the measure that you are interested in. So string edit distance ranges from zero to in this dataset, maximally 13 or 14 or something like that. So those are the string edit distance values. On the *y*-axis, you have the cumulative percentage of data points that have at least this string edit distance value.

Did I mention that this is really an unintuitive plot? It's still very useful though. It can show a variety of things that pretty much no other plots can show. So let me show you what that means. Let's take this red thing here, this little red edge here. This indicates that approximately 20% of the randomly-chosen words have string edit distances of five or less. Or, for example, this means, let's take this one here, so this is for the actual error blends, so this little edge here means that approximately 40% of the error blends, the blue line, have string edit distances of 3 and less, which means, what it boils down to is, that a curve that is lower than others, like the red one here, has higher values, and a curve that is steeper than others has lower values because it's more on the left. What does that show here? It shows that if you pick words at random, then they are least similar to each other because the red curve is ahead of all the others. It's where the high values are and here high values mean high dissimilarity. The error blends, on the other hand, those are on the left side, on the leftmost side so source words that go to error blends are most similar to each other. And then the intentional blends and the induced errors, they are somewhere in the middle. This actually shows you in a way, no disrespect, but it shows you that the people who did these error-induction experiments in the 70s' and stuff, the words they chose to use in their lab actually were more similar to intentional blends than to actual errors. Ideally, what would have happened is that the green line was on top of the blue line because that would have shown that people try to force them to make errors with words that are of such a nature that people actually *have* done errors with them. But in fact they chose the wrong types of source words, which made them more likely to be used in intentional blends. Anyway, what it shows still is that, like I said, in actual errors, the words are highly similar to each other, random words are completely dissimilar to each other, and the other two are somewhere in the middle. This graph shows the same result for the Dice coefficient and it does not come out very well, but it does show the same, the same effect.

Now if we focus on string edit distance here, so what we see, I mean this is basically what I just said. But note also that the source of error blends are pretty much globally similar to each other. So this completely neglected to look at where the similarity happens. For instance, in *channel* and *tunnel*, it's obvious similarity is at the end of both words, the *nnel* in terms of letters. That doesn't always have to be like that. In any case, in fact, if you do an analysis of where the similarity happens, then for error blends, you will see that it is across the whole word. If you do that same thing for intentional blends, you do not find that. You find that the similarity is the highest around the point where the words are cut up in the middle. In the sense like *fool* and *philosopher*, where the *l* is not just similar but is actually the same, something like that, which again is another point to show that, actually, maybe what happens to error blends is not exactly the same with intentional formation blends.

Now what about stress patterns, another way which word can be similar. So why not compare the different stress patterns in source words that enter into blends. I looked at approximately 2100 formations and coded them for their syllabic lengths, so how many syllables does each source word consist of, and what are the stress patterns? I just distinguished between stressed and unstressed, nothing more fine-grained than that. So one example here would be the blend *webinar* which obviously is from *web*, which has one syllable and is stressed, and *seminar*, which has three syllables, which are stressed-unstressed-unstressed. Or *jokelore, joke* one syllable, stressed, and *folklore*, two syllables, with this stressed pattern, *transponder* same thing, relatively straightforward. Then I counted for each of these how often you find that the words share the same stress pattern and I did that with a plot that I don't know how to call it, so I just call it crossed tabulation plot, which looks like this. So what that shows is ... You can't see it here but here it says *s, su, us* for "stressed", "unstressed ", and so on. And stressed like here and the important point to focus on is basically the quadrants in the middle. This is the case where both words have one syllable, and then here this quadrant consists of blends where the first source word has two syllables and the second has as well. And then the question becomes if both words have two syllables, do they also share the stress pattern or not? This quadrant then is the second source word has three syllables either stressed-unstressed-unstressed, unstressed-stressed-unstressed, or unstressed-unstressed-stressed, and the same thing here. If there is a pattern such that the word has a same stressed pattern then what you would hope is that the frequencies in this diagonal are high because that means that this word has this stress pattern and the corresponding other word has the same. And this type of plot, this is indicated by a blue frequencies, blue means 'something is more frequent than you would expect by chance.' Now, if you look at the diagonal, then it looks like this, and you can see that, most of the time, that is the case. There's obviously no variation when both words have one syllable. But then, when both words have two syllables, then the two blue numbers are when the two stress patterns are the same. When the two words have three syllables each, then the really high numbers are again blue, when they share the same stress pattern. When both source words have four syllables, the main diagonal usually has four words. So what happens is that if people choose words to blend, then they make sure if they have the same syllable length, they also have the same stress pattern.

Now what about semantics? Here I have only done a little bit of work, because basically semantic coding is difficult. I coded for a variety of semantic features and I will show only the results for four of them here basically because the other ones are so infrequent that it would have just made the table very huge, and it's the same type of plot again so blue numbers mean 'something is more frequent than expected by chance, it's a preferred pattern.' And what we see, for instance, is that if the formation is an authentic error blend, then the only thing that is preferred is that the two words are synonyms; that's when errors happen. And the same thing we find for induced errors because of course researchers gave people near synonyms in the hope they will conflate them. However, that is very different from what we find for intentional error blends: When people create a blend, then most of the time—then, actually synonyms, that's *dis*preferred. You don't create a blend to put two words together that pretty much mean the same thing. If anything you want to add two different things in particular co-hyponyms or things that together enter into a frame relation.

With complex clippings, something like *scifi* or *sysadmin*, something like that, there is a very strong trend for the two forms to be contractive. So likely in the case of *sysadmin*, you basically have a compound so two things that always follow each other and then you just shorten it. You don't just pick any words from something and put them together in the complex clipping—you take words that are already following each other and shorten them. Very different pattern.

Now next level. So we looked at selection of source words, what about the ordering? This part will be quick because basically we already mentioned the type of data, so we just have to look at them in a different way. We now check whether the source word lengths differ across blend types but this time, not taken all first words and all second words, but taking, of every blend, the first word and compare it to the second one. But the nice thing is that actually there is no change. In this case, the results don't make a difference. Here is one way to show this, so for the intentional blends, for example, what this shows is how many percentages of blends are there and what is the source word length difference. Here are the cases where the second source is longer, here are the cases where the first source word is longer and there is a huge difference in terms of significance there but it's not different from any of the findings that we have seen before. Pretty much the same you find for the length difference. I'm going to skip right here to the plot as well. This shows, for instance, for the error blends what the pairwise difference is of one source word to the other. This is source word one minus source word two, that might be four syllables. Here it's a frequency difference, but you can see that the difference is extremely symmetric. So for every source word that is longer in one direction, you also have cases when it's longer in the other direction. So all the results we talked about so far are the same.

So the previous conclusion, that people take more frequent word and put them at first and that happens to be the shorter one, actually stays the same. Unfortunately you will see in a moment the result do change, so we are going to talk about the third part now, namely on how source words are blended and how we might look into that. You can already see that that is the most complex part because there are so many different ways. I mean even once you have decided what to blend and what order to put them, there are so many different ways in which you can do that. Like I said before, there are essentially two opposing factors that should be considered when you analyze how words are blended, and those are recognizability and similarity. Some people have said that that is ill-defined and that it's *the one key* thing that one has to figure out—as I will argue I don't think that that's necessarily the case. But I do think one has to talk about why those might be conflicting or why they might, as I call it, counteract each other to some degree and that goes back to this similarly unintuitive statement I made earlier. So here again is the example of *Chevrolet* and *Cadillac*. Now if you only consider recognizability, then of course these two here would be the smartest blends you could make because there is no way you read this and do not know the two source words are. Because basically they are completely in there. Obviously, this contains *Cadillac* and *Chevrolet* so in terms of recognizability, these score really high. I mean you cannot miss what the source words are. However, those are not fun anymore. Then why is that? Because this blend is highly recognizable in terms of what the source words are, but the source words are not really that similar to the blend anymore. They are in there, the word *Chevrolet* has three syllables, but this word has six syllables. So while you can read it in there, there is also a lot of other stuff. So each of the source words now is not similar to the blend anymore and that of course a lot of times makes the blend not fun anymore. Something like *foolosopher*, apart from the semantic thing that is going on there, this is especially funny if you will, because the *fool* is in *foolosopher* and in fact you join the two because of the similarity—that gets lost here. So we will need a way then to quantify the similarity of source words to blends in a way that strikes a balance between these two, that sort of recognizes that this is actually highly recognizable but actually not a fun blend because of what it does to the similarity. So this is one of these cases where you can use statistical or numbering stuff but you have to devise a metric first that is sensitive to the right things but not overly sensitive. and in previous work I did that wrong. In a study published somewhere—I don't remember where actually—I argued for a similarity metric or index that was computed like this, which is the complex or mathematic way of saying this: For *channel* and *tunnel* it basically meant, six of the seven letters of *channel* make up six of the seven letters of *chunnel*. And to that you add the fact that five of the six letters of *tunnel* make five of the seven letters of the *chunnel*. Then, according to that formula that I came up with at the time that resulted in a really very high value, this is one of the highest values in the sample at the time, for this word. So at the time I thought this is good because I want a high value for *chunnel,* I mean *channel* and *tunnel* are really similar to *chunnel* so it's good that this creates a high value. However, later I realized that it's actually not that great. Because for instance the value for *brunch* and *breakfunch*, they were really similar. But even worse was this, namely these cases here, where it's kind of obvious that *Chevrolecadillac*, that that is not similar to the original word—those scored really high. So that metric was devised wrongly: It would not recognize that, it would over-prioritize recognizability but it would miss the fact that similarity can go down the drain if recognizability is way too high so something else was needed. I think a better measure is this, namely essentially the average of the two string edit distances of the two source words [to] the blend. So basically for *channel* and *tunnel* and *chunnel* that means you compute how similar is *channel* to *chunnel* and how similar is *tunnel* to *chunnel*. They are both very similar and then you take the average of the two. It's basically . . . it is also better because you are using the same measure as before, string edit distance, we've used that in another context, but now it's been adapted to work here well. If you do that for some data then you find the average string edit distance for *chunnel* is really low, and remember, this is a distance measure, so low value means high similarity. That's good, and we want that. If you apply that to *Chevrollac* and *Cadilet*, which would be good blends, I mean both are. These are blends of two three-syllable words, the resulting blend has three syllables, that's what we would want, then they're sort of in the middle, but if you apply it to the other test case, where they should say this is really bad, then that's what happens. This measure says *Chevrolecadillac*, that's a really bad one, that's highly dissimilar to what we looked at before. This measure seems to better find out what the right balance is between recognizability and similarity.

The data, a subset of all the stuff that I looked at, and this is what we find. So we have average string edit distance on the *y*-axis, the lower, the more similar. And we have the different formation types on the *x*-axis and again we can see that the error blends have the highest degree of similarity. So people who commit an error, they don't just do that with words that are highly similar. They also then blend them in a way that preserves their similarity. The induced error blends again were quite different. They were significantly different from what those researchers actually tried to simulate, so, again, maybe not the best choice of source words. Then, we have the intentional blends, which are the same as the induced error blends so maybe those researches were actually influenced by the type of intentional blends they've seen a lot. And all those are better than complex clippings and simulated words, so randomly-chosen words.

Now with that in mind, so this looks at overall similarity of the words to the blend. The other question, however, is where do people split the words up? I mean like with *Chevrolet* and *Cadillac*. How do they decide whether it's going to be *Chevrollac* or *Chedillet* or whatever else that will be? This would also be an interesting thing to look at. The question is, where is the cut-off point? Where do people decide, now I've got enough stuff from the first word, now the second word can come in? There is a variety of psycholinguistic theories that are kind of relevant to this context and have been used to varying degrees to talk about these issues. In general, activation-based models of the mental lexicon, computational search models, so models that have to do with word recognition, and then Marslen-Wilson's cohort model, a model of language comprehension. And the way this type of work often proceeds, or a notion this type of work often takes for granted, that is the uniqueness point of the word. It's the point *p* in the word at which that word can be uniquely identified from a candidate set of words. Using the example from before, whoever created *brunch* cut up the word *breakfast* after the *r* so only *br* is in there. That certainly is not the uniqueness point of *breakfast* because there's a ton of words that begin with *br*. I mean *break* and *breakfast*, and whatever else, some of them, but there are also a lot of other words. So if I give you just *b* and *r* and ask you which word is that, I mean you can guess something, but you really don't have a clue because there's too many alternatives. Now on the other hand, if I give you *breakf*, then you probably guess *breakfast*. If I give you *break* only, then you might guess *breaking, breaks*, whatever, something like that, but once I give you the *f* as well, you know what it is so in this type of approach, that would probably be the uniqueness point.

Most of the time, we don't really know what the uniqueness point is, because we cannot look into people's brains as they recognize words and see whether they guess the right thing or not, so the way that people proceed is then using what has been called the *recognition point*, which is an empirical estimate of where we think the uniqueness point probably is. That is usually defined probabilistically like this: It is the point *p* of a word where a majority of speakers, let's say 85%, can recognize the words with a high degree of confidence, let's say 80%. Not everybody has to agree that, "ok, it is going to be this word", but once enough people say, "yeah, I am pretty certain that it is this," then that's where you say the recognition point is. This exhibits a frequency effect in the sense that more frequent words are recognized faster than close competitors. This is important for something that I will show you in a moment. Now one question is that how we get these recognition points, and most of the time in the psycholinguistic literature that's being done experimentally. So you actually sit people down in a lab and give them successively more material and ask them, what is the word? what is it now if I give you a letter more, what is it now, bla bla . . . and at some point, once you collate all the results from different subjects, you know this is 80%, people wrote down the right words, so I am taking that. Now with this approach here, that would not work because I have 2000 blends that have a ton of different source words. I can't run that many subjects to get all the estimates for these uniqueness points. So I kind of need a corpus-based approach here, and this is again a way, or this is again, I hope an example, to show how corpus data can be used creatively to do a variety of things that are difficult to do otherwise.

Here is one example, you can just look at a corpus and check how many words are there that begin with something. If you go to the British National Corpus, a corpus that supposedly represents British English of the 1990s, then if I give you this letter sequence, then there is only one word that begins with that letter sequence, namely, *islamicization*, there's no other word. So basically that will be the uniqueness point where you know that can't be anything else\ if your vocabulary was that of the British National Corpus. Same thing if you do that in the CELEX database where you have phoneme transcription of words: if you give someone that *islamiciza*, sort of, IPA, then there is only one continuation, namely *tion*/[ʃən]. Now, the nice thing about this is that you can, unlike an experiment, that type of approach first gives you a uniqueness point or a recognition point, but it also gives you a much better estimate of how many competitors are there. If I sit you down in a lab, and I give you *islam*-something and then you guess what it is then—let's say everybody in this room took this, let's say there's 50 people here or something. I mean, you might write up maximally 50 different words. But of course, if you have a corpus, then you might also see the 200 other words that exist that no one thought of, so it is much more comprehensive. Plus, I also know what I call here the information distribution of that. Let's say there are 50 words that begin with *islam* or something like that. Now if all of you write these down, then I still don't know how frequent those are, how likely those are. Some words are more likely, simply because they are more frequent. If I do that in an experiment, I am not going have those data—if I look in a corpus, I can count how frequent each form is, and I know exactly what the distribution is. What one could do here is, one could approach it in a way like this. I am only going to look at the ... let's take the lower example, it doesn't matter because the curves are really similar. So this is an example using a word *agitation*, on the *x*-axis here you have successively large parts of that word, like you would in such an experiment. I give you *a*, I give you *ag*, I give you *agi*, and each time I ask you "which word do you think that is going to be?" Actually let's take the upper one, it is easier to explain. And then on the *y*-axis, you have the log of the number of different words that begin with this. And this just shows that the letter *a* is really a bad clue for the word *agitation*, because there's a ton of words begin with *a*. If I give you an *a* and ask you which word it is, I mean there's no way you would guess *agitation*: there's thousands of other words you'd rather guess before you come up with *agitation*. Then I give you the additional *g*, and that already drastically reduces the number of words that there are. *Agitation* is probably still not the first in your list, but it's much more likely. If I then also give you an *i*, again it drops considerably, and now the crucial thing is that there's way fewer types left, actually 12, I think (since this is the log), and the crucial thing to notice now is if I then give you even more stuff, you don't benefit much more anymore: Once you have the *agi*, getting another *t* and another *a*, does not help you much more than the *agi* already has. So one logic might be to say where the curve levels off, that's where the uniqueness point is.

Isn't that a nice approach? It isn't, it doesn't work. Because the database [...]. For one, that is a practical reason, the database is too large too generate all these graphs and look at them. So in my case, you have to do a plot like this for every blend, for the first and the second source word, for both phonemes and graphemes. So, if you have about 2300 formations, then you need to physically look at 9000 plots, and again I don't have the time. Second, the example that I showed you was a curve that has a really nice dent, going down like this and then leveling off. If you do that with many more examples, you will see that many don't have such a nice dent. So even a human analyst can look at this curve that has no clue where to see, ok, this is it. If it just goes down, imagine the curve just goes down like this, where do you draw the line? You can't.

But there is also a theoretical problem, and that is that we only look at the type frequency, but we don't look at something that I tried to sell you as an advantage earlier, namely, how frequent are the different types with these frequencies. So if I give you an *agi*, you know I can tell you there's 12 words left in the corpus but that's all the graph looked at. It didn't look at, is *agitation* maybe the most frequent one of these 12? So even though there are 11 other competitors, the words that I am actually looking for is so frequent that you would guess it anyway. It might just enjoy the frequency advantage. This is an example with hypothetical data. For instance, let's say I give you the first two letters of something, and let's say that's *a* and *b*. Let's assume there are only four words in the English language that start with *a* and *b*. Only four different words, and let's assume they are equally frequent. 25, 25, 25, 25 in a corpus. What that means is that *ab* is actually not a good clue for the real one that it is, because once you have the *ab*, you have no further clue as to what to guess, because everything that you might guess is equally likely. Now in another situation, what might happen is, I give you *ab*, and again there's 100 different tokens but one of them occurs 95 times and the other [three] are super rare. In that scenario, *ab* would be a really good clue because you would all guess it is probably this one. Because this you might not have ever heard of, because it is so rare. But this type of information in the previous plot didn't even consider. So that is apart from the practicality that I don't have the time to look at 9000 plots, that's also a theoretical reason—it was actually not using all the information that you can get off the corpus if you do it cleverly. So we do something else. We use what I, following a recommendation by Harald Baayen some point, called the selection point, which has another unintuitive definition, namely, this one: "The first point, after a part of word, at which that word is the most frequent word with that part." Isn't that great? It is easy, though. So the example of *agitation* again, if I give you an *a*, then I think in the CELEX database, there is this many types that start with *a*, but *agitation* is only the 595th most frequent one. Like I said earlier, there are so many words that begin with this that are more frequent, that you'd never guess *agitation*. Then I give you a *g* in addition and it still doesn't help you much because there's 137 words that begin with *ag* and again *agitation* is only the 24th most frequent one, so you are more likely to guess 23 other words before you really guess what it is. But then, once I give you *i* we're there, there are 12 types that begin with this, but *agitation* is the most frequent one. So now you could say, that's the one that I'm picking, so that why it's in blue, so that is the corpus-based estimate of this approach. It is very conservative, because it requires a complete identity, and only uses the minimum value, it doesn't use anything greater than that. So it's very conservative but it is a way that can be, with some programming skills, relatively, easily, implemented for 10s of thousands of words without problems.

Now again, the question becomes though how do we test whether whatever result we get means anything? Because if I tell you "everything gets split up like one letter more to the right from that word," then what do you compare that to? Again, we need a random baseline to see, "is that different from what happens by chance?" So, essentially what you need to check is, what are all the possible cut-off points and how far are they away from the one that was actually chosen? If you blend two words, you can split up at the first phoneme, second, third, fourth, fifth, but the speaker actually chose the third. So, how far is that away from the average positional of all the position that someone could have split it up? And that was what this approach basically does. And then, of course, then the programming work begins, because you do that for the first source what of all blends. So you don't have that for *breakfast* of *breakfast* and *lunch*. You do that for *lunch* of *breakfast* and *lunch*, but in the other direction, because from the reverse, the second source word contributes its hind part. And then you do that for complex clippings, and you do that for phonemes and graphemes. That's why you do it with a script but not manually. The result [...]. I am not going to discuss the graph in great detail, because [...]. I wanted to provide it though so you have the results, but I go to the discussion immediately. If you use that operationalization, you find that the cut-off point for blends, where people really make the split, is very close to the hypothesized ideal cut-off point. So it is really close, and much more so than an average cut-off point by chance. So it is likely at least that people go undergo some computation like "how much material do I have to provide to make it recognizable?" So the first source word is split up means exactly as this approach to predicts; the second one is a little bit earlier, so a little bit less is included. For the second source word, the blend coiners often took a little bit too much away than this approach would predict. And of course it might interesting to speculate why that might be the case.

For complex clippings, it is completely different: no similarities at all. Basically underscoring that that is probably a morphologically different process. Now, why are those results less than perfect, I mean, especially here. So why does it work really well for the first source word, but not so much for the second? One possibility might just be context. So in the case of *brunch, brunch* is really not a great example, but let's take *channel*. I mean once if you identify that the first source word, and in that case it's really tricky too because it's such a transparent blend. Let's stick to it anyway. So once if I identify the first source is probably *channel*, I mean you are already in a particular semantic domain. You have a discourse context most of the time or some situational context, plus you now know ok, whatever word I process here, which is new, and it is supposed to figure out what it is, but it got something to do with, whatever like traffic or infrastructure, something like that. So you have more clues, at the point of time you have to figure out what the second source word it is, you already have a variety of other clues. So maybe you don't need all the material. I mean the approach that I was presenting here really just uses the linguistic information—it doesn't use discourse context, it doesn't use semantic knowledge and everything, but the speaker and the hearers of course, they have that. So, one possibility is that this approach is too conservative in the sense that it doesn't take a situation or knowledge into consideration. Another one, I just want to mention this very briefly, I did not separate out of this analysis what are called neo-classical compounds. Some people would say that something like whatever *television* or something like that maybe, that's not really a blend of *telephone* and *vision* or something like that anymore: *tele* has become a morpheme already. So you can't consider this a real blend anymore, and I haven't factored those out separately here. But of course it can be done at some later stage.

So how do intentional blends happen? A not quite serious interim summary, because I want to make a very clear, this heuristic—selection, ordering, and blending—that is just a heuristic or an organizational way of talking about this. So you chose two source words, and what have we found? Well, they are similar to each other in terms of their lengths, in terms of the syllables, in terms of their stress patterns, phonemes, and graphemes, and especially so in the middle, they are often in a close semantic relationship, like a frame-semantic relationship, for example. And they of course fit what's to be said, and maybe in a funny way. And then you order then such you either, you do one or two things for intentional blends. You either leave them in a modifier-head order that they come in some expression in anyway, or you establish such a structure, or you put the shorter and more frequent word first. And then you blend them by cutting them up close to their recognition points, and fuse them in such a way that you use more of the second source word, and you do that because that maximizes the overlap, and it creates a blend that is more similar to the second source word.

Now why might be that relevant? For two reasons, one is because of this: In English at least, the modifier for adjective noun, the modifier precedes the head, the head comes second, and so it's nice, if the overall blend is more similar to its head. Second, if the blend on the whole is more similar to the head, and contains more material of source word two, then that makes source word two easier to recognize. Why is that something that needs to be dealt with? Well, remember that in a blend, the second source word will contribute its end. But you don't know where it starts. If it is a blend that you've never seen before, you read from the left to the right, and you maybe see the source word begin, I mean, you definitely see the first source word begin, but you are not sure where it ends, but you know at some point, the second one will begin. But you don't know where that is. At some point, that means the second source word, as you try to figure out what it is, you encounter it in somewhat anomalous way. I mean usually when you read text you know exactly where the new word begins, in English, and you read it from left to right. If you read an unknown blend, you kind of go through the word as you read from left to right, and you might like go like, "so this is maybe where the first source word ends", now you look at the back, "is this the other word? I don't know." Then you track back and so on so you encounter it in a very atypical way. So it helps if the second source word contributes more of itself because it's the one that is encountered in the more atypical way. And for *brunch*, that's one of the few things where *brunch* is good. I mean *lunch* is the second source word, it has five letters, and four of them, 80% of them, are in the blend. For the first source word, *breakfast* has nine letters, and only two of them go into the blend, 22%. So that is a case where that exactly works. Of course, the fact that *unch* is really helpful—how many words in English do you know that end in an *unch*? I mean *hunch, munch, lunch* obviously, but there is not a lot of others. So *unch* is pretty much a dead giveaway in a particular discourse context for what the word is. It helps to have a second word do a lot of work, basically. And then we found that the intentional blends are really different.

Now, why might you want to study that type of stuff even. Because like I said, there's a lot of conscious efforts going into this, it is not really that you find out a lot of unconscious morphological processing when you look at blends. Again, I want to point out that blends are really, in some sense, at least not that typical, because a lot of other stuff is multifactorial and probabilistic and relatively unpredictable anyway, and even the sort of free processes like blending, I mean it has to be processed by the same type linguistic system that speakers use under sort of normal constraints and it can tell us how both conscious and intentional processes interact with the subconscious processes that we use all the time. And, what is unconscious or freakish or funny or something like that at one particular point of time can also affect other types of processes at a later point of time. And as we will see in the second talk a little bit later, I mean tomorrow in the second talk of this series of three, that funny structures like this, quantitative patterns, show up in other contexts as well. And to just give you an example for this now before I discuss it more in more detail tomorrow, I want to give you one brief example here, as a preview for tomorrow. If you look at lexically fully-specified V-NP idioms, things like *kick the bucket, run the risk, lose one's cool*, and things like that, and if you look at *way*-constructions, something like *make your way to the stage, find your way to the hall, fight his way through the crowd*—Adele has written about those—then these also exhibit, interestingly enough, phonological patterning, that you normally wouldn't expect. How often do you read something about some syntactical patterns, or idioms, or arguments structure construction, and people talk about their phonology? Not a lot of times. And the thing is that, just like blending, this pattern is compatible that both involves aspects of fun (or in this case alliteration, as you will see tomorrow) but also aspects of the psycholinguistic production and comprehension. Because if you look at these structures, and again I will talk about this in more detail tomorrow, you find that they exhibit phonological similarity. So in instances of the *way*-construction, there is a particular *way*, no pun intended, that the verb of the *way*-construction is similar to the noun *way*. Or if you look at idioms like *kick the bucket* or *lose one's cool*, there is a way in which the verb is often similar to the noun phonologically. So this is actually an example of more wide-spread pattern and it may have to do with a variety of different things that we will talk about tomorrow: priming, word play and things like that would be other cases The main point to be made here is that something is in a totally different area, like syntax-lexis, here we have the same type of questions where once we know what types of patterns to look for.

Now a few comments to wrap up on what future work might look like. The first thing is a very interesting, fascinating observation, we need more data. Second, we need more comprehensive description, but then more relevantly maybe in this context, we also need more comprehensive or flexible measures of word similarity. I told you about the Dice coefficient, I told you about the string edit distance, and those can do good things, but notice also that they always only apply to one level, you apply to the phonemes, and you're are done. But the Dice coefficient doesn't do anything with additional information of, what's with the syllable structure, what's with the stress pattern, so ideally, we would have a way to handle this. Plus, we need more flexible ways to handle this, too, so for instance, in terms of what about segmental analysis? This sound is different from that one, but they have the same segmental structure, which is CVCVC. How do we include that additional information? Sometimes with regard to phonemes, the problems actually already starts at the transcription level, something like *impostinator*: do you transcribe this letter, basically do you transcribe it with a [ə], or with an [ɪ]? Depending on what you choose, it will mess all the statistics that follow. If you made that decision arbitrarily—sometimes like this, sometimes like this—you would systematically down-play all the similarity results. So how do we handle this? And then even articulatory features, this is a different phoneme, but also they are similar. A [tʃ] is more similar to a [t] than to an [l]. Can we capture this and if so, how do we do it? which will help with all sorts of comparisons that we've done.

Second, we need some better measures in other ways, so we looked at the frequency differences, so we found the first source word is usually more frequent. But as I already mentioned earlier this morning, sometimes dispersion might be a better measure. We probably need more psycholinguistic theory, because in terms of word-recognition studies that like decades of work that looks at neighborhood densities and things like that when it comes to recognizing words, and I've so far utilized very little of that. And we need more experimentation, ideally, we would be able to put people in a situation where we tell them "well, make a blend out of these words", and then we see what they do, and can we predict what they will be doing, can we predict which word they put first, and so on. And we need a ton of other things as well.

Now the one thing that I think is clear is that blends, I mean I hope that I have shown that, so blends are far from unpredictable, I mean there's a lot of things we can see about them once we operationalize things properly, do the right types of statistics, and do the right types of corpus work but we need all these types of methodological tools as well as a firm connection again to cognitive or psycholinguistic theory to back up whatever we find. Thanks.

LECTURE 3

# Quantitative Approach to Similarities in Cognitive Linguistics 2: The Phonology of Idioms

So today's talk is going to be the second one in this little series of three talks that talks about how different types of quantitative methods and corpus data can help us see structure in data that we might find interesting as cognitive linguists, and that are otherwise kind of hard to come by. Yesterday, I talked a little bit about phonological patterning in blends, and the types of word distributions we can find, and how they maybe help explain, or at least describe, how blends are formed. Today, I will talk a little bit about the internal phonological structure of idioms, or sort of patterns of similarity that we find in idioms and to what degree they might be, or seem to be at least, correlated with the semantic fixedness or versatility of particular patterns.

So what I want to start out with is this idea of a unit that Langacker has proposed in his work, and again you have seen it in a different context yesterday. So a unit is a structure that a speaker has mastered quite thoroughly to the extent that he can employ it in largely automatic faction without having to focus his attention specifically on its individual parts for their arrangement, he has no need to reflect on how to put it together. Now as we all know, in Cognitive Grammar or Construction Grammar, units can exhibit different degrees of complexity, basically running the whole range from morphemes or monomorphemic words, polymorphemic words, fully-fixed multi-word expressions, where by *fully-fixed* I mean fully-lexically specified, you have no way of changing any of the words to some other words while still using that particular expression, then partially-filled multi-word expressions, so where you can change maybe a little bit, and then basically lexically fully flexible syntactic or argument structure constructions where like the *into*-ditransitive you can put a lot of verbs, there is not really a strong bias as to what can go in there.

Now if you look at units in Cognitive Grammar, we are usually concerned with symbolic units which are basically conventionalized associations of, on the one hand, a phonological pole, so basically the form side of things, and then the semantic pole, the meaning side of things. And if you consider these two poles, then you can look at their relations in two different ways. So on the one hand, you might look at the relation between the semantic pole and the phonological pole of a particular unit, which at this point, I mean just for the sake of having a word for it, you might want to call between pole relationships, and

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_004

I will not talk about those today. I mean all that comes the big heading of arbitrariness of the sign, motivation of the links between things, iconicity and all these things that Alan Cienki has been talking about a lot.

What I want to look at today in this talk is the relationships within one pole, so what happens on one of the two sides, and there has been a lot of work in Cognitive Linguistics on semantic within-pole relationships because we look, for instance, how the different semantic pole of complex expressions, elaborate each other, or complement each other, fill each other's slots in a way, but there has been much less work on phonological within-pole relationships. But it is interesting that, sometimes, if you look at data actually, with a completely different objective, sometimes these might surface in really clear ways. And one thing that I noticed when I looked at something I will talk about later this afternoon, namely, the verb *run*, is that, for instance, there were a lot of alliterations, which I talked about in a previous paper. So if you look at idioms involving the verb "to run", then you find things like "run the risk", "run riot", "run rough hot", "run rampant" and so all of them, I mean all of the ones I gave involve alliterations, and at the time, I sort of didn't have a lot of time to spend on this, but I thought it was a funny thing to notice.

So what I am going to talk about today then, basically, is case studies at different levels of specificity of units and some additional constructions will be listed or discussed below.

So first, I want to talk about V-NP idioms which are fully lexically-filled, so cases where you actually do not have a choice: if you want to use that idiom, you have to use a particular verb and you have to use a particular head noun in the noun phrase—if you don't do that, then you basically lose the semantic integrity of this. And here are a few examples of such idioms, so "kick the bucket", meaning 'to die'; "run the risk", "to lose one's cool", 'to lose your emotional restraint' or something like that, and sometimes this can be further modified in the sense you can add an adjective or something like that, but many of them also cannot. And then I want to look at the *way*-construction in this first round, which is partially lexically-filled, because you have to use the word "way", I mean, that has to be the direct object, but you can use a variety of different verbs before it, so you can *fight your way*, you can *wend your way*, you can *weave your way*, you can *make your way*, all sorts of things, but "way" has to be there, so in that sense, it's partially lexically filled. And again, here are a few examples, so "make your way to the stage", "find your way to the hall", "find his way through the crowd", you can see clearly that the verb slot here is open for a variety of verbs, some of which are obviously motion verbs, some of which are not necessarily motion verbs per se but take on a motion reading once they're used in this construction.

Now what I want to do is I will look at how these constructions also exhibit . . ., I mean if and how so, these constructions also exhibit the type of alliteration effects that I basically stumbled across when I look at the verb "run". Then, again, and that's I guess a recurrent theme in a way, the degree to which there are alliteration effects will have to be checked against different random baselines. The logic or the argument for this basically is the same as yesterday: if you look at, let's say, V-NP idioms like those and you find that, whatever, 8% of the types exhibit alliteration, then what does that mean? I mean you have to compare to something to some baseline, to some standard of comparison to see, what is that a lot, or what is that not a lot, or what is that? So, you need some sort of baseline and, as you will see in a moment, there are different ways to compute these baselines, and in order to make sure that your findings are not an artifact of your operationalization, it's of course safest to try out several of those to make sure you got yourselves against any epiphenomena.

These random baselines are different types of statistical baselines, but there is another baseline, a more linguistically-motivated one, and that would be that you look at none conventionalized counterparts. So these idioms are highly conventionalized, you cannot change them, but they are transitive structures: a verb and a noun phrase that is a direct object, so if we find alliterations there, then one way, also to check against the baseline is to look at things that have the same structure—transitive verb and direct object—but that are not idioms. Would they have the same type of effects, and so this is what I am going to looked at here, namely at non-conventionalized transitive verb structures, and to see, do we have alliteration effects all the time we just never noticed it?

And then finally, I want to look at to what degree any effects we might find are also collocated with collocational or collostructional attraction. So *collocation* is one of these terms in corpus linguistics refers to the co-occurrence, or the preferred co-occurrence most of the time, of two particular words, so one can measure or quantify different degrees to which one word likes to occur with another word, and that is essentially what I want to do here. Collostruction is essentially the same type of concept just that it is not like collocation between words but between one word and a construction. So like in this case, it could be how much does a particular verb like to occur in this construction?

Now, the first thing one needs to talk about then is where do you count alliterations, if you suspect those are going on, then what is the database? And so for the V-NP idioms I look at the Collins Cobuild Dictionary of Idioms and I picked those idioms where the V is a full lexical verb, so it is not an auxiliary. The NP is the direct object of the V, so just a typical transitive type of structure. The verb does not take any further complements or adjuncts, and the idiom meets a particular frequency threshold so it occurs at least once per 2 million words in the corpus that they used when they compiled that dictionary. Again, here are some examples, “spill the beans”, “gain some ground”, like “reveal a secret”, “gain some ground”, “make the progress”, “get the boot”, “get fired from work”, “lend a hand”, “help out”; “bite the bullet”, “do something unpleasant to reach a goal” or something like that.

Then, the question is how to count alliterations and it seems like that might be kind of “duh”, well obviously how do you do that, but there is at least one thing I do want to mention here, so for the verb and the head noun I counted, I looked at the initial segment of the verb, so in “build bridges”, that would be the two /b/s, and “lose face” that would be the /l/ and /f/, and if there were additional content words in the direct objects, I also counted those. So for instance, as an idiom, that is, like “fight a losing battle”, so the /l/ of *losing* was also included, so since we now have three words, you can have the /f/ and /l/, “fight a losing”; you can have the /f/ and /b/, “fight a battle”; /l/ and /b/ of “losing a battle”. Same thing with “keep a straight face”, the initial segment, the/s/ sound as the initial phoneme, is included in this as well. And I mean, most of the time, this is totally uncontroversial but just to make sure, I mean I also make sure that the pronunciations are consistent, and so I checked them all with the CELEX database. And then I computed how many alliterations do I find, but then the question is again what type of baseline do you compare that to?

So there are different ways in which you can compute them. So obviously, you want to look at the word-initial phonemes. But then you can do that without regard to type and token frequencies. So, for instance, you could just say, ok, in the CELEX database, because this is such a convenient database to look at, because you got the pronunciations of tens of thousands of words in a computationally accessible way, you can just check how many phonemes are there at the beginnings of words, and then you compute the probability that two phonemes, that the phonemes of two words would be the same. So this does not take into consideration any frequencies of phonemes, so the probability of two /kj/s at the beginnings of something, I mean the /k/ sound in that case, but let’s say, a /v/, or a /w/, or something like that would all be the same. And that’s, of course, maybe not that great, because we know that phonemes are differently frequent, so there will be different biases as to how different frequently things might start with the same phoneme so a second type of baseline would be this one, namely, when you consider also how frequent phonemes are at the beginning of different word types. So there’s a lot more words in English that begin with a /t/ than with, let’s say an /l/ or something like that, so this approach would take that into consideration. And so, basically, what I did for this is I looked at the CELEX database and counted for each phoneme, how many different word types are there in English that begin with that phoneme. But then what this still misses is how frequent are these words. A word like—what did I talk about yesterday, something like *asphyxiate* is obviously not particularly frequent so maybe you want to include that information in the computation as well so the third baseline would be to also take frequencies of phonemes in the word tokens into consideration so the fact that /t/ is at the beginning of a lot of words gets augmented by the frequencies with which these words occur in a language. And so now that baseline would be based on the probability that a phoneme is the first phoneme in all the word tokens in that CELEX database, and at the time CELEX database was compiled that corpus consist of 18 million words so it is not particularly huge but it covers a lot of ground, nonetheless.

So those are three different types of statistical baselines but then what about the linguistic baseline, or the control group if you will. And so as a type of control group, I sampled randomly two transitive clauses from 170 corpus files from the ICE-GB, so that's the *International Corpus of English, the British Component*, and I took 170 files from the spoken component, and just took randomly two transitive clauses that were not idiomatic. And I counted the alliterations in the exact same way as before, so the first phoneme of the verb, the first phoneme of the head noun, and the first phonemes of any additional content words that would show up in the direct object.

Now with regard to the collocational attraction, here it is necessarily to basically, do some sort of…, to compute some sort of association measure to answer the question of whether the verb of the either the lexically fully-specified V-NP or the verb of the *way*-construction, whether they like to occur together: is there a preference that the verb and the noun are strongly attracted to each other. So for that, I retrieved the frequency of each verb, both from the idioms and from the controls, and the frequency of each head noun from the idioms and the controls, and then the co-occurrence frequency of these two words respectively in all sentences from the *British National Corpus*, the *British National Corpus* is a 100 million word corpus that is supposed to represent British English of the 1990s as I mentioned yesterday, and then I computed two different measures of collocational attraction.

And for this first case study, I use two ones I have statistical properties that are quite different from each other. So one is MI, which is the short form for Mutual Information, and the other is *t*, which is a *t*-score. And the difference of these two measures is that mutual information is extremely sensitive to low-frequency items so it tends to over-emphasize the attraction of things that are actually quite infrequent. So, for instance, mutual information is very likely to always rank proper names very highly, because in the corpus text, you might have one instance of proper name with the first and last name, but the first and last name never show up anywhere else, so it still ranks it really highly, which for some applications is nice, but for some others it's quite problematic. And so, a lot of people then also use or have suggested to also use the *t*-score because what that score does is it tends to rank collocations highly if there are highly frequent. So a proper name that is used once in a text would never be ranked really highly by a *t*-score. So the two measures capture different type of distribution characteristics and if you want to show that, again, whatever result you get is not an artifact of you choosing particularly statistic, you better make sure that you cover different ones, to cover your bases essentially. So the statistical design then for the collocational study part basically is that we are going to look at the collocational strength as computed by either one of these measures, and we look at this as a function of, so the tilde (~) here you will see a lot of those types later, this tilde here means 'as a function of'. On the one hand, the V-NP group, so is it one of the idioms or is it from the controls where we would may be not expect the strong association? And second, alliteration yes or no, so the question will be if there is an alliteration between the verb and the head noun, does that also lead or may be even facilitate in a causal sense, the high degree of attraction between things?

Now, what are the results? So these are the results concerning as it says observed and expected proportions of alliterations, so on the *y*-axis here, we have the percentage of alliterations that were observed, and the first thing to know is basically this line here, so this is the observed percentage of alliteration of types in the V-NP idioms. So all the several hundreds of idioms that I sampled, 11% of them involved alliterations, but now the question is, is that more expected by chance, because whenever you use a transitive construction you might have by accident a verb that begins with the same segment, or the same phoneme as the noun. Now, so here at the bottom with the bars, you find the different baselines. So this is the first one that doesn't care about frequencies, that just looks at how many phonemes are there at the beginning of words, and you can see that this is way smaller than the observed one. This is the one that includes type frequency, so conceptually in a way it's closest to the logic of this measure but still it is about half, meaning a little more than half as strong. Then, one that includes token frequency, this is even less than half, and then these here, this is the linguistic control group, these are transitive verb phrases that are not idiomatic compared to these which are, and again it's less than half. So there is relatively clear evidence in the idioms, I mean, the idioms have a strong preference to have alliterations whereas any of the baselines and any of the non-idiomatic ones that tendency is much much weaker. And if you do all sorts of statistical stuff on that, the results are highly significant, I mean that doesn't happen by chance.

Now, what about the collocational attraction? So is there a relationship between being idiomatic, on the one hand, or not, and alliteration or not, with when it comes to looking at collocational attraction? So again, we have two graphs here. So the two panels of the two different collocational measures, so we have the mutual information results here, and we have the *t*-score results here with the median, and then for each of the two verb groups we have two bars, so this is the controls, this is the idioms, and then within each of these groups, we have no alliteration and alliteration. And the interesting thing now basically is to compare what happens. I mean it's basically two things: one is to compare the two idioms to the two controls for each measure obviously, and then what happens within the idioms, is there a difference between alliteration, yes or no? It is quite clear I guess that when the V-NP group is an idiom, then the collocational attraction is much higher. And that is of course not really that surprising, because if it is an idiom, then there is a high chance that we use them together. But what is interesting is that also within the idioms, then, the attraction is stronger if there is an alliteration. So again, the dark bar is alliteration YES, the light bar is alliteration NO and in both cases, alliteration leads to, if you want to interpret it causally, leads to a higher degree of association such that, this one is higher than that one, and this one is much higher than that one, so it seems like indeed there is a relation between these three factors such that at least in correlative language that, alliteration, also seems to be correlated with a higher degree of collocational attraction.

Now what about the *way*-construction? Same type of set up, where do we count this? Well in this case I used data from the *British National Corpus*. Essentially, all the constructions that were found there, so this is the syntactic structure or syntactico-semantic structures, so we have a subject which is the theme that moves along some path, we have a possessive pronoun, then *way* and then we have a prepositional phrase that designates a path of the goal along which the subject, the theme, moves usually overcoming some sort of resistance. And then the constructions were retrieved by manually cleaned concordances, where we basically looked at a possessive pronoun followed by *way*, and again, yes, that meant having to read several thousands of cases to find out, well, is that actually a *way*-construction or is just a normal, I mean normal non-idiomatic way of using *way*? And to give you an idea of the number of the items we have to browse to get to this. These are the cases that are actually were *way*-constructions so, you can imagine how many more we had to discard, because they were not, but that is a corpus linguist's job. So again some examples, you have seen some before, eg, *"The British Task Force made its way across the Atlantic."* or *"The water found its way into the volcanic vent."*, whatever, cases like that.

Again then, how to count alliterations? Well, in the case of *way*-construction, note that so now basically only one slot is flexible, but the noun has to be *way*, the only thing that is flexible is the verb. So essentially, for each of construction, I only noted, I only had to note the initial phoneme of the verb and the verb slot, because the one for the object *way* is given. So "banged her way", "wound your way", obviously, I would just note this initial phoneme again, I double-checked everything against the CELEX database to make sure that is the right pronunciation, and computed the percentage of alliterations both for verb types, so how many different verb types were there, and I computed it once for each, and then average across each, and then also for the tokens, because something like "make your way", that is used a lot of times.

But again, we need a baseline, and essentially, this is the good part; All the statistical considerations are the same; so again you can just look at how many phonemes are there at the beginning of the verb, and use that as the—compute the probability of an alliteration like that. You can take into consideration the frequencies of word types, and you can take into consideration the frequency of word tokens, basically giving rise to the same three types of baseline as before. But then again, I also wanted to have a linguistic control group and so for that I looked at the British Component of the International Corpus of English again, and I looked at all the instances of *way* being used as part of a direct object, and checked them for alliterations.

Now for the collocational study, I looked at the frequency of each verb lemma in the *way*-construction. So how often do people use, I mean how frequent is "make" in a corpus, how frequent is "fight your way", "fight" in a corpus and so on, and then the frequency of the *way*-construction that I had from the corpus, because we had disambiguated all these 5,800 examples. And then I computed what is called a collexeme analysis, so that is s statistical method that a colleague and I have been talking about a lot for some years, which essentially quantifies the degree to which a particular verb, in this case, is attracted to the *way*-construction, so how much does a particular verb like to show up in this construction, and in the fifth and sixth talk I think I will talk about this much more.

And the two measures that I used here again to make sure the results generalize essentially are the most frequently-used measures in this connection, so this is the result of a statistical significance test, whose properties are not really that relevant right now—if you want to know about this more, ask maybe during the Q&A—and then a measure called Delta P, which is interesting, because it's at this point one of the very very few association measures that is directional. So it is a measure that does not just say how much do two things like to co-occur together without regard as to which comes first, but Delta P allows you to quantify how much does this thing like that thing, regardless of how the relation is the other way round. So in this case I can look at if there is a particular construction, how much does this verb like to be in there, and not just the other way around. So here then we have the statistical design of what is the association strength or collostruction strength, and how is that related to whether there is alliteration, yes or no, and again, of course with each of the two measures.

So same type of plot essentially with a slight modification in terms of the observed lines for types and tokens. The nice thing is the bottom line is the same. So if you look at the verb types that go into the *way*-construction, then you find an observed percentage of alliterations of six and a half percent. But that is higher than the two baselines, here so again there is a significant preference, basically this time we can just say, verbs in the *way*-construction, they begin with /w/ more often than you would expect by chance. Note also that this percentage is much lower than the one for the fully-lexicalized V-NP construction, remember for things like, "fight a losing battle", or "bite the bullet" or something like that; we found there was 11%. So it is more than expected by chance, but it is also less frequent than when you look at constructions where both slots are filled. This will become important in a moment.

Now if you look at tokens, it is much higher, because—and what that basically suggests is that a lot of verbs that begin with /w/ are used quite frequently, are used quite often in that construction. And so then that percentage goes up to 13%, twice as much as here. And again, this is way more than—no pun intended—than with the baseline here. And also if you look at *way* used transitively but not in the *way*-construction, then the percentage of alliteration there, so completely non-idiomatic is extremely low, just 1%. So again the bottom line is that when people use the *way*-construction, somehow the verbs that make it into the verb slot a lot, tend to stop with the same phoneme.

What about the association measures? We basically find a significant effect there too, to some extent at least, or significant difference: So if there is an alliteration of the noun *way* with the verb, then the attraction is higher than when there's not, and this one is not significant for the first measure. And for the second measure, it is marginally significant. So not great, not strong the results, but there is not a lot of different verb types that go in there, so at this point, I can only say the result is in the right direction, but it is admittedly not as strong as what I would have liked it to see. Now what does that show? So there are strong alliteration effects and they differ significantly from baselines regardless of how you compute them, so for any baseline that I chose for in each case, the observe percentage of alliterations was always higher, I mean there is no way around this. And they also differ from non-conventionalized, but otherwise analogous structures, so transitive idioms just behave differently from transitive things that are not idioms. And these are weakly but still suggestively correlated with measures of attraction, which they appear to reinforce or again, if you don't want to causal interpretation here, which they are at least correlated with.

Now, that just of course raises the next questions, namely, why is that? Does that sort of purpose, and if so, what would that purpose be? How does that happen in the first place? How does it come about and then also why alliterations? As we saw yesterday, there's a ton of ways in which words can be similar to each other, why would it be alliterations and not something else, or maybe it is something else as well. Now one possible account would be, and again, this is sort of very informal, although I would use some Cognitive Grammar terminology for this to make it a little more robust in the moment, one possible account would be that at some point, a speaker used or created an expression and because of the alliteration, that was maybe fun to use, and maybe also easier to memorize, and therefore it became reused, and maybe then at some point, often enough to become entrenched enough so that other people would pick it up.

And again, as I mentioned yesterday, when I anticipated some of the discussion of today here, this is not exactly unlike that, what happens with maybe other types of processes, like, for instance, word-formation processes like this. But still how do we account for that and, I think, that there are three basic notions that we use a lot in Cognitive Grammar that help us to talk about this in a meaningful way. One has to do with this growing recognition that similarity and analogy play important roles in language learning, language processing, language production like in priming, and things like that. Second, chunking, so the recognition that things belong together, and maybe can form, and be recognized, as one unit whose component parts you then don't have to analyze any more. And third, the notion of phonological constituents that Langacker talked about in a paper that I don't think gets quoted enough for the very cool ideas that it puts forward.

So let us look at these notions one by one. Starting with similarity, so like I said before, we know that similarity plays a huge role in many situations. One of the most important ones is probably simply just that of how similarity facilitates abstraction or generalization or schematization, if you again want to stick to Cognitive Grammar terminology. But there are also sort of more mundane, more specific ways, so for instance, novel utterances, so things you have never said before, or that a child has never said before, they are usually very similar to what has been produced previously. So we don't formulate completely new things that differ in all regards from everything else we said before but usually there is a high degree of similarity to two things we have done before linguistically. Secondly, like I said, in priming, similarity can play a huge role. So priming refers to the tendency that speakers have to reuse linguistics, either explicit lexical material or, more often, syntactic structures that people have used before, so one of the classic experiments that has shown this was done by Kay Bock, so people would read maybe a passive sentence, "The Dog was hit by the car." something like that. And then they would see a picture that showed a transitive scenario, so something that could be characterized or could be described with an active or a passive sentence, but if people had read a passive sentence before seeing that picture, then they were much more likely to describe that picture with the passive sentence too. So if they had said that "The dog was hit by the car." and then they saw something, they would be more likely to use a passive sentence then when they have read "The car hit the dog", in which case they would be more likely to use an active sentence.

Now this type of priming effect, the recycling of syntactic structures, is stronger if utterances are more similar to each other. So the more things two utterances share, the higher their degree of similarity, the stronger such priming effects are. And then as we have seen in the talk yesterday, similarity on various levels might also be conducive to forming subjective word formations, such as blends and other things. Now then the question is, but why similarity in the shape of alliterations, and if you look at at least, some psycholinguistic literature, you will find word beginnings are important points in a word, so there is one study by Noteboom (1981) that shows that word beginnings help with word recognitions more than word endings, although, one often intends to think that rhymes are so important but actually word beginnings help more. If you look at work done by Ben Bergen and a lot of other people on phonaesthemes then we find those are often located at the beginning of a word. There is a very nice study by Luca Onnis and some colleagues on artificial language learning and he pointed out—they showed—that if you have people learn an artificial model constructed language, then, that has non-adjacent syllable dependencies so that has correlations between things that are not next to each other, then people have an easier time figuring those out when those were marked with alliterations.

Now the second notion that is relevant here is that of phonological constituency. So in this paper in *Cognitive Linguistics*, Langacker distinguishes semantic and conceptual constituents and those are basically—I mean that's, basically, we talk about that all the time—so things that connect elements, fulfill valence requirements of each other, or something that semantically or conceptually elaborates some unspecified property of something else, so that's basically within-pole relationships on the semantic level, but then also phonological constituents. And those are based on, for instance, temporal contiguity, rhythmic cohesiveness, and then quoted from him, with my high-lighting, "stress, pitch level, and similarity in segmental contact". So the idea here of that paper of Langacker is basically to argue against the standard notion of constituency that is so important to formal or generative approaches grammar. So what Langacker basically argues is that what they talk about all the time, classical constituents, are just a confluence of, or arise in the situation where a semantic constituent is expressed by a phonological constituent. So if this is expressed by that then that usually comes about in the shape of classic constituent of generative or formal grammar. But that doesn't have to happen all the time. And so for example, one instance that he gives for a phonological constituent using stress is when he says—what is the example?—the example is this, and the stress assignment is important, so he says as an example "Every linguist can make generalizations, but only an **MIT** linguist can make **interesting** generalizations." Nice choice of example, but the point here obviously is that what is highlighted with the stress is *an MIT linguist* can make *interesting*. So that is not—in no syntactic theory is that a syntactic constituent. But with using stress assignment here, those two things are marked as belonging together on the conceptual level for the elaboration in this case of probably irony. And so we can use a phonological marking like this to highlight things that belong together.

Now interestingly, when he talks about semantic constituents, this is actually what he says, and again, the highlighting [...], so another kind of conceptual group is highlighted now, the semantic pole of a complex lexical item. It is well-known that idioms are often phonologically discontinuous, hence not symbolized by classical phonological constituent. And that is exactly what we are looking here, we are looking at complex lexical items, I mean "bite the dust", or "to kick the bucket", I mean that's several different words but it is used as one complex lexical item, meaning "to die".

Now if you add to these independently-made observations again from the sort of exemplar-based perspective that I talked a little bit about yesterday and that I will return to this afternoon, people have been arguing that entries sharing, phonetic and semantic features, again phonetic features of course points to articulatory similarity are highly interconnected depending upon the degree of similarity.

So the hypothesis might be something like this, so there is a certain degree of similarity that is manifested in word beginnings. And we know independently that those are salient. And then that facilitates, or leads to even, the recognition of phonological constituent. That then leads to, in terms of this exemplar-based type of approach, that leads to a higher degree of interconnectedness because of the similarity, and may give rise to chunking and ultimately constructionalization.

Now, with that in mind, what would be possible next steps? Well again kind of "duh", I mean, yeah, you need more data to look at more different types, more different tokens. Ideally one would look at other conventionalized constructions and maybe especially proverbs, and one would explore also how much they obtain similarity effects are dependent on the construction slots not being too flexible. And in order to look at that, I will add, in a moment, another construction to the mix, namely the *into*-causative. And also we might need a more comprehensive and flexible view of similarity, so right now I only looked at alliterations, but that is of course actually really impoverished, right? I mean because even in something like "keep a straight face", I only looked at the /s/ sound of "straight", although there is a whole onset cluster /str/, but I just forgot about the rest. So ideally, one would check, well does this whole theory also still work or do something if you don't just arbitrarily pick the very first phoneme, and again that relates to stuff I talked about yesterday. And then we may need some more sophisticated quantitative methodology as the data we gather become more complex, they will also actually become more intransigent in terms of how you can handle them statistically, and so we will need to basically become a little more advanced here.

In terms of theory, there are also some questions here. One is, what is the scope of this similarity effect? If you looked at Bybee's 2010 book, she recapitulates some of her earlier discussion on *strung* verbs and the degree to which how this is a class of verbs defined by similarity. On one particular page, she talks about how these verbs have similar onsets, and then a half way down the page she talks about how those also have similar rhymes. So she is very well aware of the fact, the similarity can be located again at different points in the verb. And so where does that happen? And there is a ton of possibilities obviously, I mean, it could be the first phonemes which is what I looked at so far, right? "Run the risk", the /r/ sounds are the same. It could be the similarity of the first phonemes, so in something, like "give me a break", obviously highly conventionalized expression, so that sound is not the same as that sound. But it is similar, I mean they are both plosives, they are both voiced. So the only thing they don't share is the place of articulation, but manner and voiceness, that is the same. It could be the identity of onsets. So now we are looking at the whole onsets and they might be the same, so "fly the flag", I mean /fl/ /fl/ same thing. But it could also just be the similarity of onsets, so here we have the /g/ sound as the first but there is also something else following, so the two are similar but not the same.

This could be the similarity of words as a whole. So "get the boot" in this case, both words are monosyllabic, both are CVC, consonant vowel consonant, both are voiced plosives at the beginning, and then the /t/ sound at the end, I mean there's a high degree of similarity in both parts of the same expression. And so the similarity could be all these things. Syllabifications, stress, segmental structure, it could be length, or it could be overall articulatory similarity. So here are two examples (again the coloring does not come out here for some reason) but so this is a "[ðə kæt ɪz aʊdə ðə bæg]" (The cat is out of the bag) like the proverbial saying. The two content words, "cat" and "bag", are highly similar to each other, starting with plosives, having the same vowel, ending with a plosive, or if you have something like "[meɪk hedweɪ]"(make headway), then continuant, and /ei/ in both cases. So a lot of potential sites where similarity might be studied. I am going to discuss some data on each of these, although I will not go through this in excruciating detail.

And then the question is also, how wide-spread is this? We know that similarity like this is very strong in completely-fixed proverbs and sayings, but does that also extend to things that are more flexible? Like I said, I would add the *into*-causative to the mix and I am not sure this is quite obvious by now, but I am actually using a particular cline here from, in terms of lexical fixedness: so the V-NP idioms, both slots were fixed. "Bite the bullet", you cannot put another verb in there, and you cannot put another noun in there, so two slots, but they are fixed. Now the *way*-construction had one slot that was fixed namely "way" and the other one, the verb slot was flexible. Now, in the *into*-causative, that has two open slots, so the *into*-causative is this construction here, verb, phrase that has a patient objects, and then *into* V-ing, the progressive form of some verb, so *to trick someone into believing, to bully someone into marrying, to fool someone into submitting*, or something like that, so there you have two slots but both are flexible. So we are basically looking at different degrees of lexical fixedness and later, we will check is that correlated with the degree of a similarity we find.

Then in terms of methods, how can we study such data best? We do need some control group. I mean you always need some control group for this type of stuff, so what I am going to do here is basically take the non-*way*-constructions examples and the non-idiomatic transitive examples as the linguistic control group. But another problem is that for many of the things we are going to look at now, it becomes too complicated or too annoying to just look at observed versus expected percentage differences so we will need something better than that. But the type of stuff that you would usually want to do in terms of statistics here, I mean, you run into huge problems. So ultimately what we will have is we will have a dependent variable that is probably numeric and that quantifies how high the similarity between two things. And then usually, people want to do with correlation or ANOVAs or a *t*-test something on that. With the data I will be discussing here, you cannot do any of this. And the reason for that is the data that you get when you do this type of empirical study, I mean, when you look at this type of empirical data; they violate all the assumptions of ANOVAs, of linear modeling, of *t*-tests. It is just not possible; the significant token values you get will tell you nothing, the data simply do not allow the normal type of statistics that you would want to use there. So what I will use is some statistics from the field of robust statistics which have been designed to be able to address, or to handle, data that exhibit this type of problems.

How can we look at this? Let's first look at the way measuring similarity with regard to first-phoneme and onset identity. So first-phoneme identity is what I looked at before already but now I would add the way the *into*-causative to the mix. But now I will also look at the onset identity. So "keep a straight face" that would be the /str/ sequence of "straight" would be included into the mix. So the first phonemes were identified as before, and then I checked to what degree they are identical in the two slots by basically looking at, how often does each first phoneme in the verb occur with each first phoneme in the noun. So for "run the risk", this occurs with an /r/, this occurs with an /r/, same thing; for "bite the dust", this verb starts with a /b/, the noun starts with a /d/ so that is not the same phoneme. How often does that happen? And then I computed basically a particular statistic that says, well, does it happen more or less frequently than expected. And then I compare these summary statistics for the ones that are identical and for the ones that are not, and always using the control group as a reference, so the V-NP idioms get compared to the controls. The *way*-construction gets compared to the controls, and the *into*-causative gets compared to the controls to see whether there is any further-reaching similarity effects. And then I used two particular types of statistics, the exact same nature of which is not relevant at this point, but they are robust statistics in the sense that they can handle the violations of assumptions that would normally come with these data and then I did the same for the onsets to see: are there onset identities as well?

Now, what about similarity? This is actually quite tricky question. I think we all agree that a /g/ sound is quite similar to a /b/ sound, because they are both voiced and they are both plosives. But then is the similarity between those two, is that the same degree of similarity as in a /s/ to a /f/, I mean those share characteristics, too: they are both voiceless and they are both fricatives, so they also differ with regard to manner of articulation, but we would be prepared to say, those two pairs are equally similar to each other. And then another example question would be so a /t/ and a /l/, they are similar, say, manner of articulation as a—same place of articulation, but the other two are different. How do we quantify that? It is not straightforward how you would do that, and especially, if you extend that from just a single sound to words, or to whole onsets, how do you do that? So what I took here is I took the first phonemes and the onsets of course later, too. And then the way that I checked for the similarity of the first phonemes is by using a version of something you have seen yesterday, namely a string edit distance, but one that has been modified so that it actually also considers articulatory features. So normally, if you apply a string edit distance to a /g/and a /b/ sound, you will get one, right? I mean you have to change the /g/ into the /b/ and then you are done. What that does not do is to say how different are these two sounds. So in quantitative dialectology, people have worked on this type of stuff and have devised the type of metric that can handle that differences as well. And so that will tell you to how high that degree of similarity is. And so this is then basically the implementation that I used here.

And then again, I compared the idioms to the controls, the *way*-construction to the controls, and the *into*-causative construction to the controls, using the same types of statistical tests and doing the same thing for the onsets, so there is a high degree of redundancy, here or repetition at least. So let us look at some results and I mean as you will see you get a ton of results out of this, but I promise to not go through all the slides that you have in the handbook here. So let us look at one example of the controls and the idioms and this is actually a case where having the color would be really nice, so you get the blue, so the C here is red, so the first result would be identity of the first phonemes, and again we have one of these very beautiful and very meaningful, and very unintuitive ECDF plots, cumulative distribution function plots. On the *x*-axis here, we have the Pearson residuals for identical first phonemes. What does that mean? So these values range theoretically from—infinity" to + infinity. And what they mean is if that value is 0, then something happens at chance frequency. If these values are positive, then something happens more often than expected, if those are negative then something happens less often than expected. Basically, this side means underrepresented, this means at chance level, this means overrepresented. And then we have a plot for the idiom, results and then we have a plot for the control results. And again, just like yesterday, the plot, I mean the line that leans more into the right side is characteristic for the higher numbers. And in this case, remember the higher numbers means something happens more often than expected by chance. So what that shows here is that in terms of first-phoneme identity, the idioms are on the right side so they are associated with high residuals and highly positive residuals in turn means there is more alliteration than expected. So this approach basically replicates the previous results, namely that the idioms have a significantly higher degree of alliterations than the control groups. And the significance test here is one the robust things that shows that is indeed this case.

Now, what about the similarity of first phonemes? What about "get the boot"? When /g/ and /b/ are similar, but not the same. This approach would treat those as different, it would not be able to see that actually the phonemes are quite similar. In this case, very nice, we find the exact same result, this time even highly significant. And unfortunately again, this is the polarity orientation of the axis reserved. So on the *x*-axis now, we have this string edit distance, which means that low values mean high similarity. The left side means you have to make few steps to convert one thing into the other, which means that this time around that would be hypothesized that the idioms would be on the left side of things, where things are similar to each other. And that is in fact what we find: the idioms across the whole range of the plot always are a little bit ahead of the control groups so they exhibit higher degrees of similarity. What that means is the idioms compared to the control groups—not only is there a significant preference for the phonemes to be the same, but even if they are not the same, there still a preference for them to be similar compared to control transitive structures that are not idiomatic.

Now what about identity of onsets? This graph is very interesting, it is what is called a null result, there is absolutely nothing going on. If you do a significance test here, it is not significant. So onsets, apparently, they don't have that similarity effect, so something like "keep a straight face", the /str/, that is not likely to be used in the other component word as well. If we look at similarity of onsets, however, then yes, we do find that. So first phonemes like to be identical, in the two words and similar, the onsets identity does not matter, but the idioms exhibit still more similar onsets to each other than expected by chance. And that of course is interesting, because it shows that the more precise resolution of not just looking at the first phoneme but also looking at sort of larger materials, onsets and the finer structure of similarity, the articulatory features, actually does pay off and shows something.

Now in these slides, I show you the similar types of graphs for the *way*-construction and for the *into*-causatives, but in the interest of time and not boring you to death, I will not go through all this at the same level of detail, but just go to the overall graph that chose all together.

And the thing with this, it is kind of hard to interpret, because the graph shows such a lot of information. But if you look at the first phonemes, there is at least a relatively clear pattern here and it can be summarized like this. The idioms exhibit the highest degree of similarity, followed by the *way*-construction, which is relatively similar, it is a little bit ahead, but it is relatively similar to the *into*-causative. The similarity you find there, and then you find the least degree of the similarity in the control groups. And if you look at that, then that actually, for the first phonemes, that is the perfect correlation with how semantically versatile or lexically versatile these slots are. Remember, these were the cases where both words were given, the verb and the noun, and you could not change it. And those have the highest degree of similarity. This was the case where one slot was flexible; this was the case where two slots were flexible; and this is non-idiomatic stuff where you can do whatever the hell you want. So perfect correlation in terms of I mean there's a perfect correlation such that, the more fixed the stuff is, the more suddenly you find phonological similarity in there. For the onsets, it's not quite as nice. It's a total mess, so apparently onset structure, I mean onset similarity really plays only very subordinate role which is actually really is weird, I really expected it to be, I mean, if anything may be the other way round. Because I did not expect that people would look more at the first phoneme, even if that means breaking up an onset cluster. But apparently, that is, historically at least, what happened. But overall, we can say, there are significant differences between most of these patterns, first phonemes and those of the controls.

Now, what about a broader picture of similarity? So what about not just the word beginnings? So for all the relevant words, I extracted from the CELEX database a much more precise picture of the similarity. So I took out the full transcription. This would be what you get the entire CELEX database for the verb "remember". You get it with syllabification, so first syllable, second syllable, third syllable, you get that the second syllable is stressed. So that is basically the full pronunciation annotation in here. Then, I took out the phonemic transcription without syllabification and stress, just the phonemes. I converted that also into the segmental structure, so basically now, glossing over the exact phonemes and just keeping whether something is a consonant or a vowel. Then the stress pattern, *remember*, second syllable is stressed, the syllabic length, that is 3, phoneme length, that is 8. And then I made comparisons between the different groups. So again, we have the idioms the *way*, the *into*-causative and the control groups and so for the first four: the full transcription, the phoneme transcriptions, segmental structures, and the stress pattern are used string edit distance. So how similar is one word to another in terms of phonemes, how many phonemes do you have to change around to convert one word into the other no unike what we did yesterday, when we look at the "channel" and "tunnel" and "chunnel". How different is "channel" from "chunnel"? you need to change one phoneme and then you are there. For the other two, I just computed the difference. So if one word has 3 syllables, and another word has 5 syllables, well, then they are similar to the value of two, they differ by two syllables. And then I use the same type of statistics, namely robust statistics that try to help doing ANOVA and confidence intervals to see what the results are. I am not going to show you a tons of ECDF plots and means and everything, because again, as you can imagine, if you do all these tests, you get a huge number of results but I am just going to show you an overview here, so we do get the grey, that is good. So, if you look at the full transcription, this is the order of similarity that you find. Idioms have the highest degree of similarity in terms of full transcription so, everything, phonemes, syllabification stress, everything. They are significantly more similar to each other in the *way*-constructions which are significantly more similar to each other than the control items. And then those are significantly more similar to each other than the *into*-causatives.

Now what does the grey thing mean? The grey shading here indicates the part of that sequence that is in line with the expectation that things become less similar as they become less fixed. And so the idioms are most fixed, they should be on the left of the scale, and they should be more similar than the *way*-constructions, and the *way*-constructions should have more similarity than the control but then *into* does behave as expected. *Into* should ideally be here. But three out of the four are in the right order.

If you do that for phonemic transcriptions, this is what you get. The grey thing would be probably only idiom to control, or idiom to "way", the rest does not fit. For segmental structure we get this. So again, three are in the right, but "way" is in the wrong position, "way" should have been here if it was completely well-behaved. This is what we get for the stress pattern, syllabic length, and phonemic length. The main point to be made here is that, (here I forgot to put in the grey stuff) in every one of these comparisons, at least two if not three out of four behave in the right type of order. If you test that with a permutation type of test, then that is actually a significant result. So on the whole, the correlation between the similarity and lexical fixedness for these four constructions on these different level does hold and is significantly higher than chance, the strongest results that I found for full transcription, stress pattern, and syllabic lengths.

To wrap up, so we do find significant differences between the three types of constructions on the one hand, and the non-idiomatic or non-constructionalized conventionalized patterns on the other hand, and we find that especially for identity or similarity of the first phonemes, we find some degree of that for identity and similarity of onsets and then for full transcriptions, syllabic patterns and syllabic length. So I guess I was lucky to have stumbled across "run" where it is always the first phoneme because that's what turned out to exhibit the strongest effect. And so we find that within-pole similarity of units, across morphology in syntax—, using *morphology* here because many of these idioms you can considered as complex lexical item—it is greater than expected by chance according to whatever type of statistical baseline you come up with, but also greater than what happens in non-conventionalized, but otherwise syntactically similar structures. Interestingly then, this is unlike priming, it's a very localized verb effect, I mean this happens within the single verb phrase. Priming happens like from one sentence to the next. And if you do it experimentally or corpus linguistically, you would find priming effects across like ten different sentences or something, here if you use a passive now, then ten sentences later, you are still more likely to use a passive than if you hadn't heard me use one. But this is extremely localized, within one of the same verb phrase from one content word to the next, there is a collocational attraction, but also some sort of phonological similarity force going on there, that maybe helped these things to become conventionalized.

Since we know that similarity facilitates the formation and then also the retention of these types of things, here are some other examples. We see that this holds for less flexible items as well, so the other interesting thing then here is this correlation between lexical flexibility on the one hand, and degree of phonological similarity on the other hand. So we do seem to have this type of cline with different shapes of blue indicate, a darker shades of blue indicate higher degree of fixedness, so we have completely fixed sayings or proverbs, something like this, which exhibit more similarity than idiomatic V-NPS, which allowed some modification, which was higher than for the *way*-construction, where one thought plot was flexible, which is higher than the *into*-causative, where both are flexible, which is higher for the controls, which are completely flexible. That is a very interesting observation, because one would normally not have expected this, most people do not pay a lot of attention to what happens within a particular phonological pole. And it does seem to make a sense and again I am using this quote because I basically like it a lot. That sort of different degrees of interconnectedness in a multidimensional exemplar space, these things can be manifested in different types of ways. One Question would be whether similarity like this also constrains maybe the productivity of how particular slots are filled. Would people be more or less likely to say, I have already have a noun object here so I am not completely flexible anymore as to which verb I am using? To be honest, I cannot imagine that that is the case. But on the other hand, something like that had to happen and had to be the case in order for these findings to come about in the first place.

Now, final word of caution, it is really tricky to handle these data statistically, I would like to believe that I know a little bit about that types of stuff, but some of these data really posed interesting challenges here. So, for example, in this one case, I actually don't know what to do at this point, I looked at the phonemic transcriptions of all these cases, and I computed the string edit distance that also considered articulatory features. And if you do that and you do this type of ECDF plot, you find super-strong results. So here again, we have the string edit distance at *x*-axis, so small values indicate high similarity, high values indicate low similarity, or high dissimilarity, and then we have the *way*-construction here, the idioms here, and the control group here. And so there's huge differences between these curves, highly significant in whatever direction you test. But then I thought, what does that really show? Because then it turns out that, to play around with this, I fit what is called a linear model, so basically a type of correlation or regression analysis, where I tried to predict the string edit distance on the basis of only the length of the two words. If you do that, then you get a really high correlation value, so to some extent at least, these curves are super-strongly influenced just by how long the words are. Even they are totally none articulatorily similar, if the words are of similar length, then already you get a high degree of similarity because you have to make fewer phonemic articulatory changes.

So that actually may be not what I want or is it? So then, I partialed, I made sure these values have nothing to do with the length any more. So there is a particular statistical way to partial out the effect of length out of that similarity. And then I tested that again. And then you get this, then the results are still in the same direction, and they are still highly significant. So, on the one hand, you may think, "ok, cool, it is not just length, it is articulatory similarity as well", and of course, that is good in a sense, good if you study that type of hypotheses. On the other hand, of course, it is trivial that of course our perception of how similar things are is affected by the length, I mean a two-phoneme word is less similar to a thirteen-phoneme word, then to a three-phoneme word. So I did find the effects are the same, but actually at this point I am not sure whether do I want to separate this or not. It is kind of tricky, I mean, depends on exactly what type of predictions you make, and at this point, I am not really able to say which of these I want.

So, do we want "length" to be partialed out? On the one hand, we do because we want to look at articulatory similarity; on the other hand, we don't because obviously length has something to do with what we think about how similar things are. So, it's really tricky. And another thing then is the question, well so we have these, at least, six different levels on which we can measure similarity, like full phoneme transcription with syllabification, without syllabification, phonemic length, syllable length, stress pattern, segmental pattern, plus onset, so is it possible to ultimate conflate these all? Is it possible to get one value that sets for two words how similar they are on all these things? I don't know, I have an idea, but I don't really know.

If we could do that it could be really great, because then we could do a lot of experimental testing to see how much does that actually pan out. And of course, a measure like this would be really interesting again coming back to blends, for instance, I mean suddenly, we would have one measure that says exactly how similar to source words are. It would be interesting for models of word recognition because it has an impact on neighborhood densities, so how similar are the competitors in a word recognition task. So we have a lot of implications and a lot of applications but at this point, like I said, I am actually not really sure how to handle this type of complexity. Thank you.

LECTURE 4

# Corpus-based Cognitive Semantics: Behavioral Profiles for Polysemy, Synonymy, and Antonymy

This is the third talk on the different ways in which quantitative structure, quantitative patterns, might be found in corpus-linguistic data, thereby helping to look at how particular things that otherwise are difficult to tackle can be brought to light with this type of techniques. As you can see, in this talk I want to talk about lexical semantics, basically the behavior of words in their corpus contexts in concordance lines and so on. And then this will actually be followed up by one or two talks that then talk about the same type of perspective using constructions as the point of interests. So this one will be lexical semantics, or the behavior of lexical items in context, and then afterwards I'll talk about constructions in context.

Now, what are the types of questions that you face basically when you try to do cognitive semantics? Depending on what you look at, there is a variety of different ones, and they all share something, namely that they're extremely difficult to address. When you look at polysemy for example, you often face the question of how to decide whether two usage events that you find in a corpus, or two uses of, let's say, a particular verb, or a noun or whatever, are sufficiently similar to be considered as a single sense. So if you go back in time a little bit and consider like a lot of the early polysemy analyses, I mean how many analyses of the preposition *over* do we have in English? If you go back in time to that then you see that a lot of times it really seems a lot of the senses are very similar and it was discussed for quite some time to what degree we can even say whether something is similar or different, whether something is one sense or not. And this is something that, whenever you look at any polysemy items, that is something you have to come to grips with. That, of course, in a way presupposes that you have kind of like a similarity scale on which different senses can be located so that you can say, ok these two senses are really similar in whatever multidimensional space so I consider these two uses to be one sense whereas these two things are very dissimilar. So how do you decide how far or how close to each other in terms of semantic similarity are two uses? And then once you start doing this, then you come up with kind of like a sense network, again of the early type analysis of *take* or *over* or *mother*, all these classic analyses. But then the question is that a new item that you find

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_005

in your corpus is similar to several different types of senses, and so you think where in the network of senses do I connect this to and why and how do I make this decision on a principled basis and make it replicable? Finally, there's a lot times the question of how do you determine what the prototypical sense is of a particular lexical item. Even if you just look at cognitive linguists' favorite preposition, namely *over*, then you will see that different papers have argued for different prototypes. Some sort of just being stationary *over*, some being the over-and-across sense, and they all have some sort of argument, but then it's very difficult to decide which of these is now more convincing. And some of these can be handled with corpus data, some of them are maybe a little bit more tricky to address.

Now for a different lexical relation that has not been studied that much in cognitive linguistics, I would think, similar questions arise from the notion of (near) synonymy. I put *near* here just for the sake of completeness. I mean the idea is probably no two forms have completely the exact same meaning and usage condition so even if I don't always say near-synonymy, that is what I always mean. And basically in some sense at least what you can say is, well, you actually inherit all of these problems from polysemy—you just have to take whenever it says *word* up here, you just put *sense* there, then you have the same type of problem that you need to solve. So you need to find out, for instance, what are the differences in meanings or construal or whatever between near-synonyms, and also what is the functional relationships between near-synonyms in their semantic domain. So how do particular words that mean very similar things, how do they carve up the semantic space in a way that native speakers effortlessly know what to use on what occasion.

Now if you look at how this has been done, or how these types of questions have been addressed by cognitive and other semanticists, then there's first a whole bunch of approaches were not really empirical in any rigorous sense, and this means no disrespect but I think it's fair to say, a lot of the analyses like early Lakovian type of studies, they basically assumed I mean they had some sorts of arguments, but it was very difficult to apply them in a rigorous way and what you ended up with having them basically is that even very similar uses or usage events were often considered to constitute different senses. Then we have what I will call here partially empirical approaches, and one instance of that I think could be Tyler and Evans's principled-polysemy approach because they at least make empirically very testable distributional assumptions. They introduce additional meaning components, but they have distributional features of different senses and they talk about lexical choices regarding patterns of modificational complementation, so different senses of a polysemous word would come with different complementation preferences. Now, even in the early literature already in the 1990s and before, some problems of these approaches were recognized. So for instance, in this very well-known paper by Dominiek Sandra and Sally Rice the question was raised: what is the ontological status of the proposed networks? Is that something that's in the minds of the speakers? Is that something that's only in the mind of a linguist? Do we actually mean something neurological or brainy with these types of things or it is just a representational format? And in a lot of early work that was not particularly clear. Second, not all the fine-grained distinctions that were made especially with regard to prepositional polysemy were supported by empirical data. So sorting experiments or particular types of gap-filling type of tests, they didn't support the really super fine-grained distinctions that some linguists had made. And then in a lot of cases, of course, the items or examples that were used to highlight a particular sense were somewhat artificial, they were often very decontextualized, not embedded in an actual context, and thus maybe limit the generalizability and the validity of the whole logic.

That being said, there are, of course, also a variety of empirical approaches though probably much less frequent. Some of the earliest work was done by Hans Joerg Schmid on *begin* and *start* in the *Lancaster-Oslo-Bergen Corpus*, a corpus representing British English of the 1960s, relatively small but still widely used, that is just 1 million words. And then the paper that I already mentioned Sandra and Rice's paper from 1995 in *Cognitive Linguistics* and Rice's paper on prepositional polysemy in 1996 that used a variety of experimental paradigms like sorting tasks, sorting examples of prepositional uses into categories that speakers perceive to be the same, judgment data, and sentence generation data. Raukko in a series of papers used sentence generation tasks, paraphrasing tasks, and also specifically asks subjects for prototypicality judgments, which I personally have a little bit of an issue with because it seems to be dumping the work that the linguist should be doing on a naïve subject, who of course has even less of a clue what criteria they are actually using, but that's a different story. Then, quite some interesting work done by Raymond Gibbs and his colleagues, one paper with Kishner on *just* and a paper with Teenie Matlock on *make*. They used a variety of corpus-linguistic data actually, so they looked at collocate analysis, where *R1* here means they looked at the first word to the right of the word they were interested in, basically meaning what happens immediately after the word that they were interested in. So this paper looked at *just* and then a collocate analysis of *just* at the position R1 one would mean, what is the work that *just* modifies, what follows immediately after it. They looked at colligations or syntactic patterns, what types of for instance, complementation patterns of *make* co-occurs with which senses, what types of parts of speech follow *just*, and exemplify which of *just*'s six senses that they assume, and then they correlate the senses and the syntactic patterns. And they find something that I think it's true till today, namely that there is a need to incorporate information about lexicogrammatical constructions in drawing links between different senses of a polysemous word, basically making it very clear that we do have the semantic side, but whatever we find on the semantic side, it will have syntactic or other lexical correlates that we can explore in corpus-linguistic data. And then much more recently, Dylan Glynn and some other people from the group in Leuven have been using a co-occurrence data in their correspondence type of analysis approach to highlight how different senses of different words can be brought to light on the basis on co-occurrence data.

How is this stuff being approached in corpus-based approaches that are not necessarily cognitive linguistic in nature? So all the previous stuff are basically what do people in cognitive linguistics start? Now how have people in a corpus-based semantics that do not necessarily have a cognitive perspective approached this? Lexical semantics is probably the most widely-studied area in corpus linguistics, which probably also has to do with the strong lexicographic tradition in corpus linguistics. And the main assumption has always been this, namely, distributional characteristics of an item reveal many of its semantic and functional properties and purposes, and usually you find a quote by Firth here, namely, "you may judge a word by the company it keeps" or something like that. I usually find this one much more precise, Zelig Harris: "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". So very explicit, very testable claim here unlike some of the other work that has been done. Other people have said the same, Bolinger (1968): "a difference in syntactic form always spells a difference in meaning" or, from Cruse's *Introduction to Semantics*: "the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations that it contracts with actual and potential contexts". So there's a very strong recognition that from the context in which something is used, you can infer a lot about how similar or different things are to other things.

Now this tradition has been very much alive kicking-in, especially when it comes to the research on synonymy in corpus linguistics. People have been looking at collocational information, so what are the words that occur around some other stuff? And here are some examples of work that has been done. So there's this little study of *strong* and *powerful.* So we say *tea is strong*, but not *powerful* although *strong* and *powerful* on a lot of other occasions mean pretty much the same thing. There is work on boosters and amplifiers like *absolutely,* *completely, entirely*, again very similar to each other, but when do you use what? *Big, large* and *great*, another example from Doug Biber. I myself looked at *-ic* and *-ical* adjectives, so what is the difference between *alphabetic* and *alphabetical*? If you ask native speakers, hardly anybody is able to explain what the difference is. They sometimes have a hunch, but usually they really don't know. Even worse are cases like *symmetric* and *symmetrical*. People have no idea. If you look at dictionaries, actually they don't even explain that. And often you find an entry, just a cross-reference, so you look up *symmetrical*, it says "look at *symmetric*". If you actually look at corpus data, you find that those are not the same; they do not have the same meanings. Then people have been looking at syntactic information, so what are the preferred grammatical associations of particular items? Atkins and Levin (1995) looked at *quake* and *quiver*, Biber et al. (1998) again looked at *little* versus *small*, or *begin* versus *start*. Gaëtanelle Gilquin (2003) looked at causative *get* and *have*, so like *I got my car repaired, I had my car repaired*—what is the difference? Antti Arppe, now in Edmonton looked at large number of different verbs meaning 'think' in Finnish. All of these were done based on what types of construction associations are there. Now in order to make all is a little bit more precise because what construction does, let's say a verb enter into, that is a relatively coarse-grained level, right? I mean, even a highly flexible verb can only enter so many complementation patterns. But of course a corpus has much more to offer. So in a really interesting and also, I think, under-cited study, Atkins looked at a variety of different things, so she looked at collocations in a window from seven words to the left to seven words to the right at part of speech characteristics of the headword and then she coined a term that I will be using in this talk here, namely *ID tag*, which is basically any type of collocation or colligation that correlates with a particular sense. So if a word has several senses, then basically what she is saying that each of the senses will be marked by collocation or colligation with different ID tags. So an ID tag is something that serves to identify what sense or what aspect of a meaning is highlighted in a particular usage event. In another interesting paper that provided terminology that I will use here, Hanks (1996) looked at collocations, colligations and then coined term of *sense triangulation* where the idea is if you look at collocates in different clause roles, and these things that happened in different places around, let's say a verb, they will allow you to triangulate what specific sense of a word is meant on a particular occasion. Then he used the word *behavioral profile* for something that late Dagmar Divjak and I use in a much wider sense of terms. He used it to refer to the set of complementation patterns of a word, basically saying that the semantics of a verb are determined by the totality of its complementation patterns. So here in his sense of the term *behavioral profile*, it really only refers to complementation patterns. So what type of transitive, intransitive, ditransitive, predicate nominal type of constructions—only that type of stuff. As you will see in a moment this will be extended considerably.

The problem with these approaches as I see them is that in these early papers there was very little evidence for any predictive power of these ID tags. So Sue Atkins and also Patrick Hanks, they pointed to some correlations and say we look at here is this and then this sense is meant. But there was no rigorous test of how predictive these ID tags really are and there is very little quantitative sophistication, so I mean pretty much nothing of this work went beyond percentages of occurrence. So what I want to talk about here then is the notion of *behavioral profile* that has been broadened to deal with some of these problems. So for example, the approach that I will be advocating here is more able than other approaches to take larger sets of synonyms or antonyms into consideration. So as you saw in the previous slides, most work looks at two words, maybe three, but that is a kind of it. It will be interesting if one could extend something to whole lexical fields and I think the present approach can do that. A lot of times they often focus on the base forms of the words, so for verbs like the infinitive or the third person singular or something, but do not take a variation into consideration here. And they focus on either lexical collocation or syntactic colligation but do not combine these two things. Exceptions are some that I mentioned here like Kishner and Gibbs' work or Gibbs and Matlock, who looked at both but much of the work in corpus linguistics at least says, "OK I look at collocates, that's it" and basically completely forget about syntax. And then like I said, much of this is quantitatively, not particularly sophisticated and corpus linguistics at least also lacks any type of theoretical account, basically never leaving the purely descriptive plane.

Now the way that I want to improve on this is obviously allowing larger sets of antonymous words or senses of polysemous items. The approach that I will promote here actually doesn't require the use of different word forms but it really encourages in a way that you will recognize in a moment and it includes a huge range of distributional characteristics, so not just collocations, not just colligations, but a large number of things. And to anticipate the question, yes, those have to be manually coded. And it includes statistical analysis that goes beyond just reporting of percentages although as you will see the approach is actually based on percentage type of analysis.

So what are the four steps? Step 1 is kind of boring and self-evident. It involves the retrieval of, hopefully, ideally, a large and representative sample of all the instances of a word's lemma from corpus in their context, that's usually at least in the complete utterance or sentence. So if as I did at some point in time, if you are interested in the verb *run*, that means you run a concordance of all forms of *run* that you can get out of a particular corpus.

Second step is the annotation, so far largely manual analysis and annotation of many properties of the use of the word forms, and that's where the notion of ID tag comes in. So the characteristics that I included or that can be included in such an analysis basically range, I mean, run the whole range from morphological characteristics, syntactic, semantic, and whatever else you might want to include. This can be dozens of characteristics, ok?

Step three is basically generating a set of co-occurrence tables that indicate how often each single one of these features occurs with a particular sense, if you are looking at a polysemous word, or how often each of these things occurs with one of the set of synonyms. So if you look at five synonymous words, then this step would mean to say for each of the features that you annotate, how frequent is it with each of the five synonyms? If you look at a word that has nine senses, then this step means to cross-tabulate for each of the nine senses how often it occurs with these things, and I will exemplify this in a moment. And the last step is of course the statistical evaluation of the data that you get, because the amount of data you get out of there is insane, I mean, you can not possibly look at this like a spread sheet and make any sense of that. You will need some statistical tool that help you make sense of this huge amount of data. And so one might start with descriptive summary statistics to see how often what happens, one might add on top of that a few correlation methods, and I will give you an example for this, but the main application so far has been this, namely, a cluster analysis, to identify in this huge data set, to identify things that are more similar to each other than they are to everything else on the data.

So this is the overview of applications that I want to give you here. So we will talk about polysemy, near synonymy, and near synonymy compared to antonymy. So basically senses, lexical items that are really similar to each other, and then lexical items that have both similarity and oppositeness relations to each other. And then each of these will hopefully exemplify in an insightful way, a particular way of how these things can be made sense of in terms of a statistical analysis.

So let's begin with what actually was one of the first studies in this area, and then the study will also serve to highlight the sequence of four steps. For this study, I looked at 815 instances of the verb lemma *run* from two corpora, again the British Component of *the International Corpus of English* and then the good old *Brown Corpus* representing American English from the 1960s. These were annotated for 252 ID tag levels. So each of the 815 instances was looked at and was coded for stuff. Those included morphological ID tags. So for the verb lemma *run*, for each form I coded what tense is it, what aspect is it, which voice is it in? It included syntactic ID tags, so is *run* used transitively, intransitively, ditransitively, or whatever else, complex transitive? Is it used in a main or subordinate clause? Is it used in a simple sentence, a complex sentence, a coordinate sentence, all sorts of things . . ., and it should be clear, I guess, that there was no expectation actually that the uses of *run* differ in terms of whether they go into a main clause or a subordinate clause. I mean there is no prior hypothesis: this is exploratory. Then, semantic ID tags, that is the greatest pain ever, because if you've ever tried to do semantic annotation on corpus data, something as noisy and authentic as corpus data, you will know why. So basically what was done here, subjects, objects, complements, they all were annotated for semantic roles that they exemplified. And then even worse, more torturing, in this case at least I had to come up with the label for the senses of *run*. So I had to basically decide, this is a different sense at this point, this is a different sense at this point, and then gave it names, some of which maybe more telling than others. Generally speaking, this will be what I will argue is the prototypes, so the 'fast pedestrian motion' type of sense that we all know. But then there are things like 'be a candidate' that will be exemplified by *he was running for office*, like *he was running to be President* or whatever. 'Operate', that's like *he is running a forklift* or something like that. 'Manage', that's *he is running a pizza store*, so there is a lot of different senses like that. *Run* actually really is highly extremely polysemous. *Water is running down the drain. Tears are running down their eyes.* I mean all sorts of things that only have to do with motion can be used with *run*. Then I looked at lexical collocates in the same clause and all of this was entered into a table like this, so you would have a column that says what sense a use of *run* was attested in and then this is the form that was actually observed, which is the base form, and so this would be a case, let me take the third one here. The third one represents the case where *running* was used which is progressive, it was used in the sense of 'manage', so like *he was running a pizza store*. This verb was used in a subordinate clause with a countable subject but not an object, and it was used intransitively, and so on. So a huge number of data points that went to that analysis. And then at some point at least later, I wrote a script that takes this table as input, and converts it into a different table, namely, one that looks like this, so now the senses that were sort of in the different rows, here in this column, now have individual columns. And then each of the ID tags such as clause type, main versus subordinate and so on are in the rows, and the crucial thing now that this table does is is the following. It allows you to compare the percentages of the senses for each ID tag. So for instance here you can see that the 'spatial extension' sense of *run* is attested 60% of the time in subordinate clauses and 40% of the time in main clauses. That is very different from the 'manage' sense of *run*, which is attested 74% of the time in subordinate clauses and only 26% of the time here, which again is completely different from what this one does. This one, 'fast motion' is actually most of the time in main clauses, and less so in subordinate clauses. And so crucially then these percentages within an ID tag add up to one. I mean you can do the math yourself, this plus this plus this, I mean that's one, 100%. So within each characteristic that you annotate those add up to 100%. Again you can see that, obviously, there are differences. Although verb forms here does not seem to differ a lot—these are always in the 90s or something, these are always much less—but here there are huge differences. And it is this type of difference that the analysis will want to exploit. What in our approach then is a behavioral profile is not what Hanks said—so Hanks would just say like the complementation patterns—what we call a behavior profile is this series of numbers from the top of the table to the bottom of the table, that's what we call a *behavioral profile*, a series of percentages that specify for each sense or for each word how it is used. From this table you can see 'fast motion' is used 94% of the time in intransitives, 5.9% are mono-transitive, not anywhere else. This sense is used this much time in this and this much time in that. So, this column of numbers, that's the behavior profile in our approach.

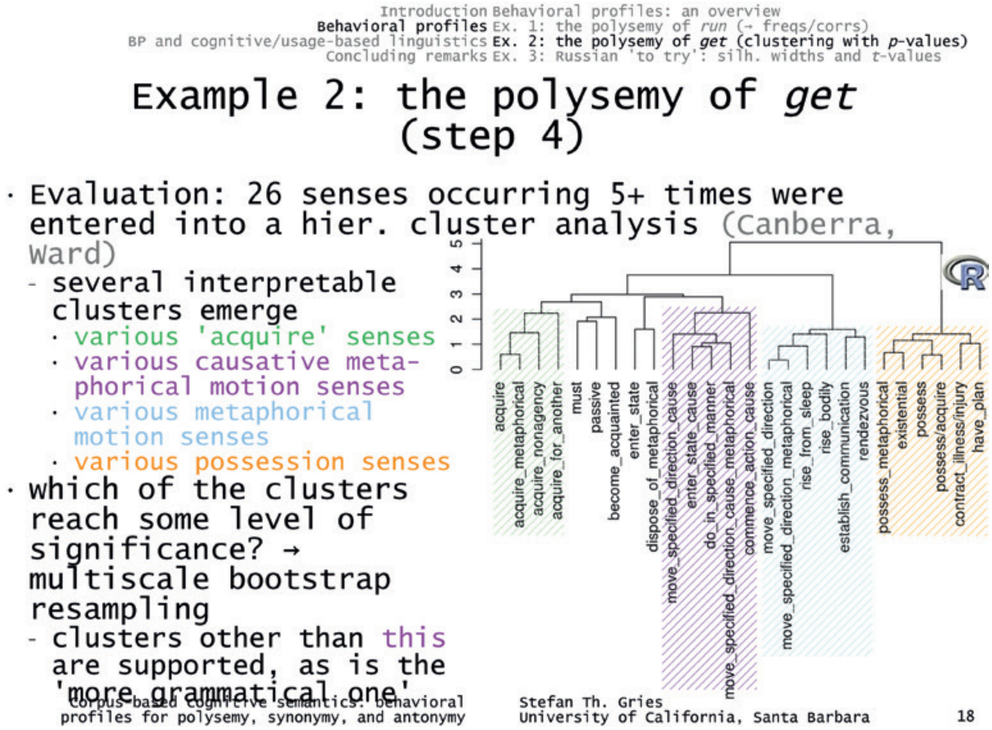
Now how can this be used? First example here is not yet highly statistical. To some extent, you can use it to answer this question, namely, which senses that you assume are in the data would you want to lump into one larger sense and which of them you want to keep separate and be split? There is a very nice example by Bill Croft in an early paper, where he makes a point that can be answered, or examined, with behavioral profiles. So he looks at the polysemy of the word *eat*, of the verb *eat*. Here has these two examples: *Jack ate lunch with Jill*, where *eat* means 'dine', and then *Jack ate a pizza with Jill*, where *eat* means 'consume', with a comitative, attested to it. Then, the interesting thing is that this one here, *Jack ate lunch with Jill*, I mean, you find that in corpus data, so this is not a particular dish, but a particular type of meal. This one, perfectly grammatical, this is not wrong, but people don't use it, they don't say it, so the sense of *eat*, that is 'consume' *with Jill* or something like that construction, is not attested in corpus data, and he uses that to say, apparently one of the two senses of *eat* takes a particular type of structure that the other does not, so there is a distributional difference and that means these are two different senses.

Now for our *run* analysis, that would mean we can lump the following senses, look here, we have, these are actual examples from the corpus. *And we run back to the car*, so we have *run* meaning something like, I mean running fast 'fast pedestrian motion' with a GOAL argument, so where does it go? Where does the running go? We have a case, and more cases of course, and we also have a case like this with a SOURCE argument. And we have a case where SOURCE and GOAL are combined, even though you might be tempted to say "those might be different senses", the fact they share the same distribution, they allow for the same types of arguments, would force you to conclude, I mean if you take that approach, that actually those are the same, those are not different senses. 'Running somewhere' and 'running from somewhere'; if it is this type of sense, those are the same. That also means, however, you would have to split several senses that seem to mean something like 'escape'. So for example, again these are real examples from the corpus, *If Adelia had felt about someone as Henrietta felt about Charles, would she have run away with him*? When I looked at this, my first hunch was to code it as something like this, 'To move away to engage in a romantic relationship'. But there is also something like this, *He wanted to know if my father has beaten me or my mother had run away*. So here, this is 'to move away from something that is undesirable'. Obviously, these two senses are similar: they mean 'to move away'. But the motivation is different. And interestingly, you do not find cases where you have both the comitative and the source in the same sentence. So the distributional difference then would force you to say, those are actually different senses. They have something in common, but they also have something not in common, which is the complementation patterns and so I have to keep them separate.

The second question, you might want to answer with this type of approach: where to connect senses in a network. Once you decide these two 'escape' senses are similar, where do you connect them into the sense network of *run*? So you might want to connect them to 'fast pedestrian motion', which is the prototype, because prototypically, 'to escape from something' would be by running quickly, so why should it not go with the prototype? But then you can escape something without that being fast. You might escape from somewhere by boarding a ship, and then that ship goes across the Atlantic or something, which is not exactly fast compared to other modes of transportation. Would you connect it only to 'motion' or would you connect it to 'motion that is fast but not pedestrian'? How to say, I mean how do you decide that? One way to do this would be by computing correlations between the different senses. That would mean to go back to that slide here. That would mean you compute the correlation of these numbers with these, and with these, and with these. So you compute all pair-wise correlations and check, "ok, this sense, is it more similar to this, or it is more similar to that?" You just let the similarity relation decide where to connect a sense in a network. Interestingly enough, if you do that with these data, you get a very clear result. On the whole, the senses are all really closely, relatively positively closely related to each other. First validation finding, if you look at the most dissimilar senses according to this method, you find that these two senses are most dissimilar: *Their cups were already running over without us* and *he ran his eye along the roof copings*. Those are really very different senses of *run*. Now the two most similar senses are [in fact I did not know this the time I did this], but in fact is 'escape' and the prototype. So if anything, then on the basis of the corpus data at least, this is the type of connection you should make.

Second example, looking at another super polysemous verb in English, namely the verb *to get*. This is work done with a former students of ours. We looked at only a small sample because obviously *get* is so frequent you can not possibly annotate all instances with all these features. So we looked at 600 examples. We annotated them for 54 different criteria, and again those were all from all sorts of different types of analysis: morphological features, syntactic characteristics, semantics ID tags, where we coded for the abstractness or concreteness of the sense. Here is a variety of senses that we found, so obviously that is the 'acquire' sense, 'stable possession', 'movement in a particular direction', 'entering into a state', like *I got dizzy*, that is not receiving, that is 'I am feeling dizzy'. Then we did this conversion to get this co-occurrence table, where again every column was one sense of *get* and then we have the numbers that said how often does that sense occur with particular things. Then we took the 26 senses that had at least a frequency of five because otherwise the data become too sparse. Enter them into a hierarchical cluster analysis, which is a statistical techniques that yields this type of output, so it yields a tree diagram structure, and the way to read this structure is basically such that things that are merged close together at the bottom of things—something like this here, this is the first thing that is merged on the vertical axis—those are highly similar. Long vertical lines like this [pointing at the graph on the slide]
indicates that something is dissimilar from the rest. That means that this orange group of senses here, those are highly similar to each other, but they're quite different from all the rest. Same thing here. So this light blue, I don't know whether you can see that, but this light blue set of senses, those are internally coherent, but they're really different from the rest. Same here.

What are the senses that we find? The green stuff here on the left, those are a variety of 'acquire' type of senses, so the typical 'acquire' sense, but then also 'acquiring something metaphorically' or without agency or for someone else. There is a variety of 'causative metaphorical motion' senses, which are nicely brought out by this purple cluster. A bunch of metaphorical senses and various senses having to do with 'possession', which are the orange one here: 'possession metaphorical', 'possession acquire', and 'have a plan'. We actually find that there is a lot of structures in this data set although not all the features



Introduction | Behavioral profiles: an overview
Behavioral profiles | Ex. 1: the polysemy of *run* (→ freqs/corrs)
BP and cognitive/usage-based linguistics | **Ex. 2: the polysemy of *get* (clustering with *p*-values)**
Concluding remarks | Ex. 3: Russian 'to try': silh. widths and *t*-values

## Example 2: the polysemy of *get* (step 4)

- Evaluation: 26 senses occurring 5+ times were entered into a hier. cluster analysis (Canberra, Ward)
  - several interpretable clusters emerge
    - various 'acquire' senses
    - various causative metaphorical motion senses
    - various metaphorical motion senses
    - various possession senses
- which of the clusters reach some level of significance? → multiscale bootstrap resampling
  - clusters other than this are supported, as is the 'more grammatical one'

Corpus-based cognitive semantics: behavioral profiles for polysemy, synonymy, and antonymy | Stefan Th. Gries University of California, Santa Barbara | 18

we included in the beginning, you would have reasonably expected to help with this. So we did not feed any senses basically into the analysis, but we get a really good structural picture out of that on the basis of the quantitative algorithm. Which of these clusters reach some level of significance? There is a way in which cluster analysis can be exploited with significance tests as well and as you can see, there are at least three or four clusters that are highly coherent, so coherent in and of themselves, that the analysis says "you get to treat that differently from the rest". I am going to interpret the result with great details. This is basically just a way to show you that this is in fact a possibility that you can do more analytical statistics on top of things.

An example involving near-synonymy, using examples from Russian here. We had about 1600 instances of nine verbs that mean 'to try' from various different corpus sources. They were annotated for 87 features, or ID tags. It is pretty much always the same: we try to cast as wide enough as possible when it comes to include the potentially predictive features, converted this into a co-occurrence table, and then did a cluster analysis on this with the same statistical features and this is the result we got. We have now not different senses but different words. We have nine different verbs and the analysis strongly suggests that they come in three different clusters that share distributional behavior. Using a follow-up statistic, this number three—there is a way to test whether they are really three not something else—and that supported that type of result, that type of interpretation. This is the result, and it basically says when you assume three clusters, then you have the highest degree of discriminatory power in the analysis, so that's what we did. But then the question becomes "what do they represent?" I mean "what do they indicate?" There are two ways in which one can relatively easily follow up on this. One is you can compute the so-called *t*-statistic that tells you, which are the things you annotated all the 87 features, which of these are highly characteristics for one of these clusters? So, for instance, is this a cluster where the subject is often the human, or animated? Is this maybe a cluster where the subject is usually inanimate or something like that? You can just read out the number, you do not have to guess, you do not have to intuit something—you just get it out of the data. Secondly within the cluster, you can compute how these are different? So these three verbs share a lot with each other, which is why they form one group, but of course, they are not the same; they still have differences from each other and that is what that approach will tell you. So when you do that, for instance, you find for this first cluster, [*poryvat'sja norovit' silit'sja*], whereever the stress goes, you find they have inanimate subjects, usually they go with physical motion verbs, and they refer to actions that are uncontrollable, but often repeated, especially maybe because you do not succeed, right? Remember those verbs mean 'to try'. So if you do not succeed with something, especially because they are uncontrollable than you may have to try a lot of times. Then, with these verbs we have inanimate subjects, often metaphorical and physical motion, and a high degree of vainness, so a lot of times you do not succeed. And then with this one you have animate subjects, but a lot of times they do not undertake that attempt denoted by *try* voluntarily but they are exhorted to do it so they perform an adverb use of intensity. If you compare that to the previous lexical graphic works by Arpresjan and others, then you will find that that is compatible with, but also goes beyond, what these previous studies have shown.

Another example, involving this time a contrastive study, so we are looking at data from two languages at the same time—English versus Russian—with phrasal verbs. So, in English, we looked at examples of *begin* and *start* and, in Russian, we looked at examples of the corresponding translations, both aspectual partners, and in that case we had to use a journalistic data. We annotated them again with several dozen features from the same range of expressions and also, again as usual, for the sense they have. Now the other nice thing you can do with these behavioral profile vectors you can see where things are maximally different. Let me actually go back to that slide to show you that. If you want to compare these two senses, then the behavioral profile, all that is offers these factors of numbers. But what you can do for each of the features that you annotated, intransitive, mono-transitive, copula, other, main, subordinate, . . . you can compute pair-wise difference. You can say this sense is attested within a main clauses, 40.4% of the time, this one is attested in main clause 40.8% of the time, which means, that is pretty much the same value. There is very little difference. With regard to these two senses, there is a huge difference. This sense is attested in intransitive clauses 94.1% of the time. This minus that. That is a difference of 94%. So there is a huge difference between these two senses. That is of course because 'manage' is a transitive use and 'running' like that is not. You can compute for every, if you want to compare two things, you just compute the difference of the two and then sort by the difference so that the biggest difference will be on top and you see what to look at first. Since the annotation here is for features that can be compared across different languages, we can do exactly that, even if we have examples or languages that are otherwise really different. Here is a comparison within a language, we took the verb *begin* and we took the verb *start* and then we listed all the ID tag features, depending on how big the difference is. You can look at what types of result do you get there. For instance, where are the big differences here? This is difficult to read. I will show you a better representation of this in a moment. A lot of differences have to do with well what is it that is begun. We have a sort of intellectual or mental things, linguistic unit, which is like a word or something like that, or we have some sort of action, something like that whereas if you look at *start*, what's begun, a lot of times there is actually nothing that mentions, so it is used intransitively, but then we have what is begun 'action', so 'general action and communication', from just sorting the differences, you can see what the particular word is used more often than the competitor that you are comparing it to. So *begin*, compared to *start*, is used in main clauses with progressives a lot when nothing that expresses begins or something abstract whereas *start* is mostly used transitively and a lot of time in subordinate clauses and when someone that is human starts something, especially an action or a type of communication. So again you do not have to leave the differences to your intuition or something, you just take them out of the corpus data, and see where they are.

You can do the same for the Russian data and in the interest of time, I will keep that brief. Then you can even do a cluster analysis of the words in different languages to see what is the word in one language that behaves most similarly to some other words in some other language?, which obviously can be very useful. In this case, you find that *begin* is most similar to *nacinat'/nacat'*, and *start* is somewhat similar to *stat'* in, for instance, the sense that all these verbs prefer zero or more abstract beginners, and that these two prefer past tense and similar beginnees, namely, action, communication, or mental activities. I think again in the interest of time, I am gonna to leave it at that, but if you have a question about this, you have the slides, please do let me know.

Second to last example: comparisons between the first language and a second language variant. We are going to look at very briefly at the English modal verbs *can* and *may*. We had a large amount of data on *can* and *may* from native speakers. So the LOCNESS corpus is a native-speaker essay-writing corpus from L1 speakers of English Then, we used the French component of the International Corpus of Learner English. This is essays written by French learners of English and we have a large number of examples from there. And then we took a French corpus; French native corpus where people were using the verb *pouvoir* which can be used where you use *can* and *may* all the time and annotated those for in this case a smaller set of features but again from the whole range of variables one might be interested in.

Then we entered the data into the logistics regression model, with the idea is that we try from the things we annotated we try to predict: would a speaker use *can* or *may* on this occasion and hopefully also why? We eliminated predictors that have no statistical significant effect on the choice and then found it came up with the final model with extremely high classification accuracy, so in 99% of the cases, the statistical model was able to predict whether the speaker would use *can* or *may*. These are some of the results that we had. For instance, on the *y*-axis, we have the predicted probability or percentage of *can*. We find for instance with regard to sentence type—interrogative vs. declarative—that makes a big difference for the probability of people using *can* or *may*. What else could we take? So we looked at the verb type, for instance: *states* are different from *achievements, process*, or *accomplishments*, so the Vendlerian semantic classes a big difference there and if you combine all these things together, then you arrive at a really high predicative accuracy. Interestingly then, you can also look at interactions that tell you something about where learners have difficulties. Where do learners still not get it completely how to use *can* or *may* compared to a native speaker? There is one example here, for instance, that involves the presence or absence of negation in the data. We have the learner data, interlanguage, and we have the native language data and then we distinguish between affirmative clauses and negative clauses in both cases. As you can see here, all speakers in general prefer *can* in negated clauses, so the black bar here says people will use *can* and that is very high when the clause is negated. But the L2 speakers do so much more strongly. One way to explain this might be if you are non-native speaker then using negation is an additional thing to process, it's something else to keep track off, which might then result in the speakers, say, not planning of course, but amounts to something like, people are already in the process of overload which makes it more likely they fall back on the more frequently default verb, which is *can*. There is some work by other people that goes in that same direction. I may be going to skip the second example in the interest of time. Let me just move to the next example, because I do want to say something about theoretical implications.

The last example here involves both synonymy and antonymy, which is an interesting test case for a variety of reasons. One of them is that antonyms are actually tricky in the sense that, everybody says "well ok, synonyms are words that are similar to each other, and antonyms they usually denote the oppositeness". However, if something is antonymy or something else, that really means that the two words are still very similar to each other, they just have opposite values on one dimension. If you take *hard* and *soft*, I mean, those are opposites, they still share a lot of things, namely that you apply them to concrete objects, that you make that judgment on basis of how the surface feels or something like that, so in a variety of ways, the words are still very similar. It's just on the one dimension that they refer to, they inhabit of the opposite end of scale. So the question that is can the behavioral profile approach actually handle this because, behavioral profiles, the idea is that similarity in meaning is reflected in the similarity in distribution. Now if synonyms and antonyms are actually still both relatively similar to each other, will that approach be able to see what are synonyms, what are antonyms? In a study to find out whether that works or not, we looked at the size adjectives in English. We looked at *big, large* and *great* which are near synonyms to some extent at least, but then we also looked at the opposite *little, small*, and *tiny*. We annotated a huge number of ID tag levels, many of which were collocational in nature. We entered into the script that I wrote that actually does everything. If you want add that script, the script tends to input of a particular corpus table and does all the rest. It computes a cluster analysis, and this is the result that we got. It's very interesting because a lot of things that have been argued in the literature are really confirmed here. For instance, the fact that *big* and *little* are antonyms. The opposite of *big* would not normally be *small* but *little*, depending on what you apply it to. The fact that the *large* and *small* are antonyms, that is reflected in the analysis. We have a nice synonym cluster here, namely *smallest* and *tiny* which mean pretty much exactly the same thing. Then we have comparative forms and we note these are morphologically marked, I mean comparatives, or superlatives, that behave alike. Then we have *large* and *smaller*, and *greater*. So there is a lot of structure here that make sense if you compare it to previous work. Those are the cases that I just mentioned. Here we only have base forms, then we have the red comparatives. This whole structure, and then the orange stuff here on the right that I mentioned. Again the question though is what are the differences then between *little* and *big*, and what are the differences between *large* and *big*? This is now the better representation from this plot before where you could not read it very well because it's on the side. For every annotated feature, we just note the difference and every difference that is not in this white area in the middle is significantly different from zero. If you want to know what the difference is between *big* and *little* in terms of where they occur for instance: well, *big* occurs more often in independent clauses, *little* more in the main clauses, *big* is used more often with abstract things, with organizations, institutions, groups, and communication, whereas *little* is overwhelmingly used for concrete, animate, and human. If we apply that same logic to near synonyms *large, big*, we can see what a lot of people would instinctively recognize, and I mean especially as native speakers, namely that *large* is often used for quantities, you say *a large amount* and not *a big amount; a large number*, and not *a big number* something like that. Again the crucial thing you can just read that out of the distributional data, you know, immediately what you can discard and you know immediately the numbers, the distributional characteristics that highlight what is important when you want to compare these two and on the next slide—I am not going to discuss it again—on the next slide, you basically have this type of summary that tells you what the corpus data show with regard to how these words are used.

Now interim summary: so the approach shows if you look at the size adjective before, you needed several different studies before. So they confirm, for instance, the results of a rating study: Deese in the 1960s did a rating study where he basically found exactly this: *smallest, tiny* which is something that came right out of the cluster analysis. The dendrograms show the *big, little, and large* and small are canonical antonyms, which is what most studies have shown There's an experimentally reported tendency that the subject like to respond to a particular stimulus with the response that has the same morphological form. This is what we saw on the cluster analysis where morphologically identical—morphologically identically-marked items, had showed up in the same cluster. I did not talk about this here, but we made sure this is not just artifact maybe syntactical features or semantic features you find with both or with either of the two.

Now, why does this approach work? Why is these numbers of features that seem to be completely unrelated still give rise to so much structure and how can we interpret this? In a way, I think we can answer this with something that I mentioned in the very first talk, namely that this type of approach taps into the information that we represent in our minds when we process all these exemplars and basically store them in a multidimensional model. It supports the view that corpus linguistics is not a theory but a method but that it is one that is very closely related to psycholinguistic assumptions of exemplar models. What are the assumptions of these models? Again, every time you hear or produce something you place a memory trace of this piece of information into a multidimensional space or network. These are quotes that you have seen before but that I wanted to reintroduce you because they really make clear what this is all about. And again to just make that very clear: The distributional characteristics involve all sorts of different things, phonetics, phonology, prosody, morphological, syntactic characteristics, semantics, discourse pragmatics, and any other type of co-occurrence information that you might find relevant, which is one of the reasons why we cast our net so wide and annotated all these things no one ever said would be important. If you add them all to the mix you still get the really high predication accuracy.

Learning, memory, and categorization in such an approach basically mean—and you've seen a more primitive version of this type of graph before—is that basically you enter a new element into a network depending on where the central tendency of that network is. If you have some dimension *x* of something and some dimension *y* of something, now you enter a new point, let's say here, exactly here, then this point where I am pointing right now is closer to the red center than to the blue center so mostly likely it will be made a category—I mean if the context allows for that—this point here will be made a category of the red center, which means a red triangle will be added, which means this means will move a little bit in that direction, and the whole system gets updated for when the next information comes in. Here is another example, this little *x* then for instance now that will probably be made part of the blue group what slightly changes the balance here. But also that is important to notice the blue point now is already much denser and richer than the red one, so the additional one more data point here will have less of an impact on the blue data than it would on the red data, because the blue cloud is already denser, more established, and more entrenched.

We have extremely rich memory representation of events, but we do not remember everything. One, because we might not notice. Second, memories might decay over time, that is technical way of saying 'we forget stuff'. And, they may be subject to generalization, abstraction and schematization and all sorts of other things. That means the fact that our memory is not perfect actually helps, in that sense, it's a good thing: It helps identify the typical contexts, so as Ellis said, "abstraction is an automatic consequences of aggregate activation of high-frequency exemplars". Again you have seen this quote before, but I think it fits in this context. I'll skip that one.

How does that relate to corpus linguistics? Obviously, it is a corpus-linguistic method, but there is also a connection to psycholinguistic work, like from the early 1990s, from the early 2000s, so Miller and Charles, when they looked at synonymy, antonymy and so on, they notice how we a lot of times perceive associations between things that occur together in the same structure, and a behavioral profile approach taps exactly into that: It involves relative frequencies, percentages, how often do things happen? and then with additional statistics analysis like discriminant analysis, like the logistic regression approach that I showed you, it involves extremely fine-grained information on what occurs together with something else. And I think I will just leave at this and thank you for your attention.

LECTURE 5

# Constructions and Their Semantics/Behavior: Collostructional Analysis

Last talk was about the behavior of lexical items in corpora and today I want to talk about is the behavior of constructions in corpora, and the specific type of analysis. Well, *analysis* is maybe too deep of a word here, the specific type of an approach that helps find out distributional patterns of constructions and words go into them and maybe what that reveals, on the one hand, about the constructions themselves, and on the other hand, what they might reveal about the types of mechanisms that the cognitive systems uses, or pays attention to, when constructions are learned, when they are used, when they are put into a particular context.

Again there is a slight mini-series of talk, so this talk here is the first of the three talks we have to do with constructions and behavior and corpora on the one hand, and that will then segue into the, basically I think, the ways that we should look at the frequency in corpus data and a lot of times I will use constructions and their behavior in corpora as an example for that. So this is the first part of that three-talk mini series.

Now, if we look at corpus linguistics approaches again, first maybe without regard as to what actually has been done in cognitive linguistics, then corpus linguistics in general has been a really rapid going methodological tool. But a lot of the studies that have been done in corpus linguistics, as I've mentioned before, have particularly focused on lexical semantics and again I think that is in part of the fact that corpus linguistics has a strong lexicographic tradition.

So what people have done a lot is looking at key words in context, concordances of lexical items or of lemmas, basically looking at a word in its context as it occurred with a few words to the right and a few words to the left or maybe the whole sentence, and then analyze it in terms of what happened there. And a lot of times that was then facilitated/ornamented by collocational approaches, so by some sort of statistic, for instance that helps quantify which words prefer to show up in a particular context.

Now this is not to say that people in corpus linguistics haven't looked at syntax and on the whole, I think it's a fact to say that lexis-based approach has been heavily favored and part of it of course has to do with the fact that it is easier to look at words than to look at syntactic constructions, because most corpora or a lot of corpora are, if anything, part-of-speech annotated,

DOI 10.1163/9789004336223_006

so you can look for adjectives or nouns or verbs, but it is much more difficult to recover syntactic constructions, given the flexibility that they can come in.

Now, this talk then, obviously, I want to talk about more, what happens if we look at syntactic constructions and the types of patterns we find there. Now, given the overall orientation of this event, it's probably not a big surprise that I will not assume a strict divide between lexis on the one hand and syntax on the other, so I'll basically follow approaches here that in corpus linguistics would be, for instance, called as Pattern Grammar where people have argued that lexical items and syntactic patterns should not a priori be distinguished in a qualitative way and cognitive linguistics or construction grammar terms in more theoretical approaches. So I will not assume that syntax and lexis are qualitatively different and even though the main focus here will be on a syntactic constructions, of course, obviously, I'll also have to involve some lexical aspects as well.

In the previous talk I've already shown you I think two quotes that highlight that patterns in corpus linguistics or in Pattern Grammar and constructions in Construction Grammar or Cognitive Grammar, are actually two really very similar items. So here again is the probably the best known quote by Hunston and Francis (2000:37) on what a pattern is: So it can be identified if "a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it." If you compare that to Goldberg's most recent definition of construction in her 2006 book: "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist". In addition, now the frequency argument, "patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency." So we have here these confluences are different, basically, how different theoretical approaches or methodological approaches basically recognize the same type of but just give them different names.

Now the one thing that I want to add here is a quantitative perspective because in a lot of cases, especially in Pattern Grammar, the quantitative sophistication with which people have looked at patterns or constructions has not been that high. I think that is important especially because I want to make it very clear that corpus linguistics, that I consider it as a pure distribution of discipline. What I mean by that is that corpora per se do not contain any of the things that linguists are interested in most of the time, so a corpus per se does not contain meaning or functions or concepts or anything like that. What corpora contain is essentially information on frequencies of occurrence or relative frequencies of occurrence or percentages, so corpora contain something *x* many times, and *x* might be zero, i.e. it is not a test in corpus or it might be greater than zero. A corpus contains information on dispersion, so any one element you are interested in might occur a number of times in particular parts of corpus or of corpora, or things might occur at particular distances from each other. And corpora contain information on relative frequencies of co-occurrence, so here we have frequency of occurrence, so how often does something happen in the corpus", and here we have frequency of co-occurrence, so how often does something occur in the corpus with something else or in the close proximity or something else? And again *collocation* would be one term that we use to refer to lexical co-occurrence, and *colligation* and later in this talk maybe *collostruction* we'll use as a term to identify co-occurrence of a word with a particular grammatical pattern. And there are all sorts of derivatives of these pieces of information, so the corpus linguistic approach of key words, which I will not talk about here a lot because it's not that relevant from cognitive linguistic perspective, but it's a well established method in corpus linguistics, it is essentially just a particular way of combining, sort of comparing frequencies of occurrence across different corpora. But again the point to be made is that, I mean none of these things in and of itself are meanings or functions or concepts or anything like that. Whenever you want to talk about meaning as a corpus linguist, you have to find a way to operationalize that in terms of frequencies of occurrence, dispersion, and frequencies of co-occurrence. For example, we've seen yesterday in the behavioral profile talk how you might approach the meaning of things by counting relative frequency, how often does a particular sense co-occur with a particular word, a particular pattern, and so on? But it is only by virtue of this extra step that you actually arrive at that something that linguists are most interested in, and not just at numbers.

Now if you look at recent work in Pattern Grammar and Construction Grammar and related types of approaches, and if you look at how they study patterns and constructions, they have so far relied mainly on raw frequencies of occurrence or co-occurrence, and then usually sorted frequency lists have been eyeballed. So, what people would do is to look at a particular construction, that construction might have be identified from some corpus either words without syntactic annotation, and then they would count how often particular words would occur in that corpus, sort the items that occur in that corpus by their frequency, and then talk about usually the most frequent items of course. And that has happened in a lot of different types of applications, so for instance people would look at a particular lexical item or a node word, and then they would look at structure/POS (part of speech)-sensitive collocates, so you might be interested in a particular adjective and you will look at the nouns that it modifies, or you might be interested in a particular verb and then you will take a look at the subject nouns that it takes. Secondly people have looked at grammatically-defined frames especially with part-of-speech tagged of corpora, I'm looking at an adjective if it modifies a subsequent noun, I'll look at *N+N* pairs, I'll look at *N+P+N*—I mean noun-preposition-noun sequences—and so on. Or they might have looked at what has been called in the early 1990s *collocational frameworks*, so cases, basically tri-grams, so three words sequences, but where the first and the third are fixed and then the middle one can vary. In a framework like this (*a+N+of*) a lot of times you will find is that the end here is some sort of quantificational noun like *a pair of* or *a couple of*, things like that. And then they have been looking at colligations / grammatical patterns / constructions and here are a few examples. Some of those might be lexically partially-specified, so the *into*-causative that you've seen in the second talk in the series here, lexemes like *to V+someone+into+V-ing*, that might be an example, *V+from+V-ing*, so to prevent someone from doing something, for example, the *way*-construction, the "waiting to happen" construction, so "this is an accident waiting to happen" which would be a prototypical in instantiation. Or they look at lexically-unspecified patterns of constructions, so *V-NP* will be an example of transitive construction, the *V+NP+NP*, an example of ditransitive construction and so on. I'm sure you're all familiar with these, and some studies have looked at these types of things.

Now it will not come as a great surprise that I think some of these approaches have been somewhat problematic for lack of statistical sophistication and I want to spend a little bit of time on why that might be an issue. One problem is that if you do that, you might miss out on important information. Here is one particular example from Hunston & Francis' book *Pattern Grammar*, where they look at, for instance, the *into*-causatives, so the *V+into+V-ing* construction, *to force someone into doing something, to bully someone* or *trick someone into doing something*. And they look at corpus data obviously but they do it in a variety of ways that undercuts the efficiency of that approach. So one thing is that this construction, as you've seen before, has two open slots, the first verb and second verb. Now in their discussion of this construction they actually focus on only one of the two, namely on the first verb slot. So basically completely forgetting, I mean not forgetting but *discounting* for the purpose of that chapter, the second slot. Secondly, they try to talk about the semantics of this construction, and they make the—I think—correct observation that the construction has something to do, or communicates 'some sort of forcefulness or even coercion'. And they attributed that of course, again correctly, to the verbs that go into the first slot, so if you do the concordance of that construction then you'll find a lot of verbs of forcefulness, I mean like *to force, to bully, to bludgeon, to beat* or all sorts of things like that. So obviously yes, these verbs endow the construction with that sort of meaning, but funnily enough, while they talk about how these verbs indicate some sort of forcefulness or coercion, the verbs *force* and *coerce* are actually absent from the list of verbs that they discuss in the data themselves. So they arrive at the conclusion but the noun comprehensive analysis they choose actually doesn't fully warrant that. In fact, the verb whose occurrence is highest in their list is *talk*, so in their data this slot here is most frequently occupied by that verb, and that is not a verb of forcefulness or coercion. That's one thing and the second thing is that *talk* is a very frequent verb in general, that means it shows up everywhere. So if a particular verb is very frequent, then you wouldn't be surprised that it shows up in a lot of different contexts. I mean none of you would be surprised to see that the verb *do* or the verb *get* shows up somewhere with a reasonable frequency. Yes, because it shows up frequently everywhere. So we probably need a way to basically normalize the occurrence of a word in a particular construction against its overall frequency to make sure that something doesn't just happen because that word happens to be everywhere.

And one big body of work that has been trying to look into this in corpus linguistics at least is that it has to do with collocational strength, an association measures of the type that I have been mentioned before here. Now collocations have to do with lexical co-occurrence, so where does one word occur given that another one word that occurs. But of course once you assume that there is no divide between syntax and lexis, there is really nothing out there that prohibits the application of these collocational measures also to co-occurrence of lexical items and syntactic items. And this is then what I want talk about here. And that of course means if that is the case then we hopefully can improve on previous results by not just using raw frequencies but some sort of association measure or normalization to get better results. And so what I first want to do basically is talk about this one such approach that a colleague and I basically have pioneered and then discuss some of its advantages and areas of application.

The first type of analysis, the sort of conceptually simplest one, is what we called *collexeme analysis*, and so what it does is it quantifies and puts a number on the degree to which a particular word and a particular construction are attracted to each other. Specifically, how much does a particular word, compared to many other words, like to occur in one slot of a particular pattern or construction that can be filled? The example I want to discuss here first is that of the *as*-predicative and the way the argument here will proceed that I want to first talk about in more traditional terms namely by discussing how a pure frequency-based approach will do this; then I hope to convince you that that is not good enough and then suggest an alternative and exemplify how I think that is superior. So the *as*-predicative is this construction here, it's a VP construction so we have a verb, here *saw*, and then the direct object of that verb, then the word *as*, and then some complement of that word *as* which can be a noun phrase as in this example or it can be a verb phrase as in that example, so this has the variety of different realizations this can take on. One question might be how is that used and is there a particular meaning that it has? So one possibility might be to say we try to and infer the meaning of this construction on the basis of what happens in this slot, what types of verbs would go in there and how often do they do that? Now if you do that on the basis of, let's say, the British Component of the International Corpus of English, which is a great corpus to do this with because it's syntactically parsed. And then you sort the verbs according to the frequency with which they occur in this particular slot, and this is the result that you get. So in that corpus the verb *see* is the most frequent word in this construction, followed by *described*, followed by *regard, know, use, treat*, and so on.

Now as I've indicated already it shouldn't take a long time to figure out that extraction, maybe not a greatest result ever, because these frequencies now have not been normalized according to how frequent is that verb everywhere else. This is the information that I've added now so this part of the table, this column and this column you've seen on the previous slide, but now I have added the overall frequency of this verb, of this verb, this verb and so on in the corpus. And so something like *see* is the most frequent verb in the *as*-predicative, but it is also a really frequent verb in general. So it is not really a big surprise that that number is high (verbs' number), among other reasons that number might just be high because this is high—the verb in general is quite frequent. So we all want to have some sort of normalization here so that verbs don't just show up because they have a high frequency here.

Now some people might say, in fact they have said, that doesn't really matter just because the verb occurs in that construction frequently, that is still already sufficient to establish an association between the verb on the one hand and the construction it appears in on the other hand. So it basically boils down to be an empirical questions: these people will say this number here is the only thing that counts, and other people including myself and some other folks of course would say "well, you need to normalize this number against that number in some way", that is gonna be important. So it is an empirical test, a question and we want to test some point and we ran an experiment. What we did is we chose sets of verbs that occur in the *as*-predicative frequently, and another sets of verbs that occur in *as*-predicative infrequently from that frequency list. And then for each verb we created an active and a passive sentence fragment, basically because this construction, the *as*-predicative is really frequent in the passive, so we didn't want to just give people active constructions because that would not have been very representative of how that pattern is usually used. So people will get something like this, *The biographer depicted the young philosopher* ______, or the passive version *The young philosopher was depicted* _____. And then the question was how do they continue that sentence, so 64 subjects were given such types of fragments and then they were asked to complete with a grammatically correct English sentence. And all sorts of usual experimental controls of that stuff were done, so they were filler items with different verbs and different constructions, the order of presentation was randomized, I mean the usual staff.

Now what are the results? We've got 493 responses that we could unambiguously code as an *as*-predicative or something else. And then when the verb was from the high frequency group, then the proportion of *as*-predicatives was indeed higher. So just like people who only think that frequency is important would have predicted—that is indeed what we find in some sense at least. Here in this bar plot, we have percentage of *as*-predicatives in the sentence completion by the subjects, and then we have that for the high frequency verbs in the *as*-predicatives and low frequency verb in the *as*-predicatives, and we can see that if the verb was frequent in the *as*-predicative then the *as*-predicative completion rate was nearly 5% higher, that difference between the two. Now happily for us, that difference, however, is not statistically significant, so it can happen by chance, which of course is kind of what we wanted, but the conclusion from that result then is that the frequency of occurrence in this particular experiment at least was not a good predictor of what subjects would do.

So that's already an empirical problem, but there are other voices which frequency occurrence might be a little bit problematic. So one is that if you deal with small frequency items, then percentages can very quickly inflate the results and in corpora, low frequency items actually always constitute the majority of tokens. So in this particular case we have—I mean stupidest example but still—the verb *re-elect* in this corpus occurs one time at all in the corpus and that one time as in *as*-predicative, so if you summarize that with a percentage then obviously you have to say 100%. On the other hand it is obvious 100% that, when the sample size is one, that's maybe the smartest way of summarizing that. Second, if you just report observed frequencies you actually don't know whether the number is greater than expected by chance or less than expected by chance. So in this particular case we have the verb *think of*, which occurs in this corpus 206 times and 6 times of these is in *as*-predicative. Now is that a lot? Is that more than you would think by chance or is that less? It's actually impossible to say on the basis of this table: six is much less than 206 but then there is a ton of other constructions. So it's really hard to evaluate that number. With a more statistical approach, we can see that that number is actually higher than expected by chance. One would construct a table like this where you take these numbers here and make them a part of the table that also contains the overall corpus size in constructions and the frequency of the *as*-predicative. And then you can compute that the observed frequency is in fact higher than expected, and we can see that it is even dependent on the corpus sample size. If the *as*-predicative was much more frequent than it really is, then this would be less than expected, but given the frequency of *has*, this is actually more than expected, so *think of* is a verb that actually likes to occur in the *as*-predicative. You just cannot see from this numbers alone, not possible. So the question then becomes, how do we know what the expected frequency is, how do we compute it and how do we find out whether that difference is statistically significant?

Here is another example to highlight how this is computed, this uses the ditransitive construction, so V+NP+NP, and we are going to use the verb *bring* as an example. In a particular corpus set, there were 461 instances of *bring*, 7 of those occurred in the ditransitive and that's basically 1.518%. The question now is: is that more or less than expected? And the way this is then computed is like this, namely you again take the overall frequency of the ditransitive into consideration and corpus size, and then you just do this little multiplication here, you multiply this, times this, and divided by that $\left(\frac{1035}{138664}\times 461 \approx 3.44\right)$, and then you'll get a number which is 3.44, and that is the expected frequency. Obviously, seven is larger than that so actually *bring* likes to occur in the ditransitive. So these ratios basically have to be compared to each other. And then, like I said, since observed is greater than expected, the ditransitive attracts *bring*—the question then becomes is that different large enough to be significant or is it such a small difference that can happen by chance we shouldn't attribute much importance to it. This example at least is not large enough, just about not. So we would have to say *bring* likes to occur in ditransitive but actually not to an extent that is maybe not random.

Now there are a lot of association measures out there—like I mentioned the other day the largest survey that I've seen compares more than 80, one particular useful one though, we think is the so-called Fisher-Yates exact test. It is a test that for the table like this for instance yields a *p*-value that can be expressed like this. I mean just for those who know—the Fisher-Yates exact test is a test that is conceptually similar at least in terms of how it is applied to the type of data to what is applied to the Chi-square test, but it works, I mean computationally, it works very differently, but that's kind of what it tries to do. And then since numbers like these are often a little bit cumbersome to express the way we have usually transformed that number by taking the negative log to the base of 10, so something like this:—$\log_{10} p_{\text{Fisher-Yates exact test}}$= 3.209) becomes 3.2. And then in this case, verbs are sorted according to that measure, which we called *collexeme strength* as an association measure and the question then becomes what does that make a difference, I mean is that actually a particular . . ., I mean does that help us in accounting for what a construction does? Now obviously doing this is kind of tedious, if you have a lot of different verbs then this approach requires that you do cross-tabulation like this for every verb. So if you have 500 verbs occurring in one construction then you need to do 500 tables like this, get this, do the log and whatever and obviously that's not fun. So I wrote a script that I might be able to show you how it works at the end with a little demo if that works here. This is basically a screenshot of how that works. So the nice thing about the script is that is interactive. Basically you don't have to know anything, you just have to answer some questions. So the script asks you which kind of analysis do you want to perform, and then you choose a number; and then it moves on and it asks what is the word you want to investigate, so you say *as*-predicative; then it asks you for the corpus size, you put that in; it asks you for the frequency of construction you put that in; then it asks you which association measure you want to use and the default is the Fisher-Yates exact test, so you choose that; then it asks how you want the output sorted and how many decimals you want where you output supposed to go like in a text file or not, and so on. So you just answer all these questions and then you get a result, if you have it on the screen then it looks like this. It's basically a table that lists all the verbs in the data, you have to provide the frequencies of these verbs. It lists how often the word is in the corpus, how often it is in the construction, what the expected frequency is, whether the verb is attracted to the construction or not, and then here, in the last row, this is the collocational measure, so you can see it's sorted from high to low so that you can focus on the top, let's say 10, 20, whatever ones, very quickly.

Now if you look at this output here we can see there is still, I mean *regard* is now on the top, not *see*. But *see* is still No. 3, and *describe, know*, and *use* and stuff like that are still very much at the top. So it seems like . . ., well, is that really worth of it, I mean, that's a lot of work for actually something that looks so similar. So we should validate this approach and in this experiment that I've showed you a partial result of before, we actually did not just include frequency as a predictor but also collexeme-strength groups. So this is the result that you've seen before, high vs. low, but we also had a frequency, but we also had a collexeme-strength, high vs. low. So that we could cross the four, and cross the two levels of each variable to arrive at four different combinations. If you then do a statistical analysis of whether a frequency on its own or collexeme-strength has a higher predictive power then you'll actually find that the collexeme-strength approach or the numbers that result from that, is nearly four times as strong as that frequency. If you look at this visually it looks like this. As before, we have frequency of the verb in the as-predicative on the *x*-axis. So high vs. low, and the last graph you saw had these two bar plots here, this is the line that represented the previous results you saw and this is the line that you represented before. And we saw last time that difference was not significant. Now, if we add the collexeme-strength results with blue (the color in the graph) for high and orange for low collexeme-strength, then you'll see that there is a huge difference between the two such that when collexeme-strengths is high then you get a lot of *as*-predicatives. When it is low, you don't; and you'll find that result regard this of the frequency. And that is the nearly-four-times' difference that I mentioned before. If you analyze this with a logistics regression and everything then you get a ton of results, and the most the important one is this point that I want to focus on is just the comparison of collexeme-strength and frequency and, as you can see, in a multifactorial context, the effect of frequency, comes with a *p*-value of nearly 10%, so it is not statistically significant according to the usual standard of point of five where is the one that collexeme-strength is highly significant. And then, effect size if you want to have a look at, that would be possible, too.

So we had first encouraging results from this sentence-completion task, but then the paper only appeared five years later but actually we did it in the following year. We did a study that tested something like that as well with a self-paced reading task. So, subjects were given sentences to read that involve an *as*-predicative or not and they differ with regard to the verbs and their predictiveness for the *as*-predicative and the question basically was, does a verb that is strongly attracted to the *as*-predicative make…, I mean does it ready people's attention for an *as*-predicative so that when that actually happens they are faster, because the verb already made it clear to them all then I'm going to get this *as* and then a noun phrase or something like that. And it turns out that it nearly did, so what we got is that again comparing just collexeme-strength and frequency for the moment, as you can see collexeme-strength is marginally significant so it misses 5% threshold just about, which of course is not so nice. However, the result for frequency is completely insignificant with a *p*-value of nearly 30%. So the difference is in the right, I mean right for us, in the sense of predicted direction and there is again the difference of effect size here of roughly 1 to 3. If anything this result is in the right direction. If we did one-tailed test, which of course we could because we have a one-tailed hypothesis, then this even would be significant. I'm reporting the overall, the normal two-tailed statistics here, because that is software and normally returns.

But if we did the one-tailed test to which we were entitled in this case, given our hypothesis, the result would in fact be significant.

So what are the advantages of this approach? It is in line with the assumption that there is no strict dichotomy between syntax and lexis, because the statistical approach that is used is one that has been developed for lexical items only, but it still produces useful results. It is descriptively more adequate, that's at least what we think because it downtones or dampens the effect of words of frequent everywhere and provides the direction of an effect by which I mean that it says whether something more or less frequent than expected. It provides an impression of the robustness of the statistics so we can see whether something is significant or not, and at least in the two experiments that I've reported on here so far, it has a larger degree of predictive power compared to raw frequency alone.

Now as a next step, we wanted to take this a little further and developed essentially two extensions of it. The first of these is called *distinctive collexeme analysis*. And the difference here is that collexeme analysis looks at one word in one construction, the distinctive collexeme analysis looks at one word in a slot of two or more constructions. So for instance, are there verbs that distinguish between the *will*-future and the *going to*-future, are there some verbs that are more likely to use the future when it is *will* or some with *going*, are there verbs that are said more likely to use *actives* or *passives*? And then you can extend that to three patterns or more, so you can see are there verbs that prefer the *active* or *be-passive* or over the *get-passive*? Can we make the statements about what these constructions do and like and mean depending on the corpus distribution?

The nice thing about it that it's actually all the same. It is all based on the same type of two by two co-occurrence table. The only change is that we still have the verb here in the first row, but the change now is that we don't just have one construction here—now we have one construction and then other—but now we have one construction compared to the other construction. So this table compares the frequency of *award* in the ditransitive as opposed to the *to*-dative, so here *he awarded him a prize, he awarded a prize to him*, which of the two is preferred? And the second change is that this frequency now is not the overall frequency of constructions in the corpus—it's the frequency of these two constructions. This many ditransitives in the corpus, this many *to*-datives in the corpus, so that is the total. If you run the type of statistical test that I've talked about before on this table, you actually get a significant result, which means that the verb *award* actually likes to occur in the ditransitive, and one can actually already see that from this distribution because, look, the *to*-dative is nearly two times as frequent as the ditransitive in that corpus. However, the verb *award* is two times more frequent in the less frequent construction than in the more frequent construction, so you would expect this to be the other way around, but it is not. If you log, this you'll get a value 1.58, whatever, and again you can sort that according to the ranking you get for many different verbs that might be called *distinctive collexeme strength*.

Here is a different example for this and again using *bring* in the ditransitive, so we're now comparing the frequency of *bring* in the ditransitive to the one in the *to*-dative relatively to how frequent the two constructions are. The expected frequency is computed as before, so you take this number times that number, divided by the overall corpus number, you get a value of 48. That means if everything was happening by chance, then you would expect this number to be 48, but it is much much smaller. So you would expect to see a lot of cases of *bring* in ditransitive, but you don't get to see them. That means *bring* repels the ditransitive, it doesn't like to occur in it. And, by implication, then it likes the *to*-dative—is that significant? Yes, it is. If you compute that, then normal statistical software would output this type of result, which, if you sort of rewrite it, means something like that: so a super small number and this number, I mean this way of writing the number now tells you why we use the log because this is a little bit easier process. So it is a very strong result.

If we apply that to the dative alternation as a whole, it returns all the verbs that are characteristic for these patterns, so we can sort them for each construction. For the ditransitive, we get *give* and then *tell* and then *show* and then *offer* and *allow* and *cost* and so on. And if you've ever read any of, let's say Adele Goldberg's analyses of the ditransitive, then you'll see those are exact the verbs, I mean, that have a lot to do with transfer and exemplify all the sense instantiations, the chief POS related for the ditransitive: change possession, communication, satisfaction condition, all of that type of stuff. And also many of these, so there's been this iconicity account for the dative alternation that said the ditransitive construction implies a closer physical proximity between the agent and the recipient. I mean, typically, if you give someone to something else, then you're standing next to each other, and it's a gesture like this so you're close. If you look at the *to*-dative, then the most strongly attracted verb is *bring*, and *bring* means a greater distance, I mean, if I'm standing next to you and I'll give you the mic I mean I don't *bring* it to you, I *give* it to you. If you're at the back of the room, I'll take the mic to you there, I mean then that's *bringing* and that is correlated with a greater physical distance. Now you might wonder so then, what is *play* doing here, for that you have to know what the corpus is, so these instances of *play* or not playing a game, but that is from soccer, so playing a ball across some distance, so even there we have a very nice pattern of physical distance reflected in the collexemes.

If you apply this to a variety of other case studies, a lot of other phenomena, you can come up with a large number of case studies that show really interesting things. Here are some examples of what people have used this approach for, so Gilquin (2006) looked at periphrastic causatives in English, Wulff (2006) looked at the difference between *go and V* and *go V*, and I think in a later study she (2008) looked at *try and V* and *try V*, so you can say *I am going to try and do this this afternoon*, or you can say *I am going to try do this this afternoon*—what is the difference between these two, do they attract different verbs? And it turns out they do to some extent. A colleague and I (2007) looked at the modification of hedges, so what type of things happen after *kind of* and after *sort of*. People have used this type of stuff to look at language change, so over time, over the last three centuries what were the different verbs that go with *shall*, and how have those changed? Hilpert's (2008) book on *Germanic Future Constructions* applies this to a whole host of data. There's really a lot of different types of applications, and in one study that I will come back to very briefly later you can even try at least to use that if you have a lot of data, to use that to look at maybe some sort of cultural differences. So we looked at the *into*-causative again but comparing British and American English, and you actually find quite some different patterns there although you would think, I mean, you probably would not expect those.

Other applications would be in second foreign language acquisition or learning, so we—Wulff and I—looked at *to*-dative alternation as used by German learners of English, so to what degree do they get that ditransitives and *to*-datives have different verbal preferences? We also looked at *to* verses *ing*-complementation, so *he started to smoke* and *he started smoking*, do these two patterns attract different verbs, and again they do. And there's implications for psycholinguistic processing, so there is a very nice study by a former M.A. student of mine, Daniel Wiechmann (2008), who looked at which of these measures that you might use, correlate with psycholinguistic data. We found cases of these verb preferences having an effect on the strength of priming. Priming, remember I told you this on the first day I think, it's the tendency to reuse syntactic constructions. Now obviously if a verb has a particularly strong preference for a particular construction, then it takes more to make a speaker not use that verb with that construction, because that's what their preference would be. So if we look at syntactic priming, this type of effect in fact can be quite illuminating as well.

Second extension of third type of analysis is what we called *covarying collexeme analysis*, so these are the two ones we looked at so far, and then covarying collexeme analysis quantifies the attraction of one word in one slot of a construction to another word in another slot in the same construction. So, you have one construction and it has two open slots, like the *into*-causative or something like that and the question is how does what happens in one of those slots, co-vary, or is determined or affected by what happens in another slot?

So here is an example of the *into*-causative, we are looking at the frequency of *blackmail into accepting*. In this little sample, a case here, the real data was much larger, let's pretend we have 200 *into*-causatives, and in the first verb slot there was *blackmail*, which occurred 51 times all together in the first slot, so in this toy example it's quite frequent, un realistically so. *Accepting* occurs 22 times in the second slot, so this is the first slot and this is the second slot. And there are 14 cases where the two occur together, so where the construction actually was *blackmail someone into accepting*. And the question then becomes is that greater than expected by chance or not, is that what we would expect if there was no semantic or other preference out there, or is that something that reflects something? So we compute the expected frequency, this times that, divided by that, we get 5.6. So that frequency is actually nearly 3 times high as you would think by chance. That means *blackmail* likes to occur with *accepting* in this construction. And is that significant? Yes, it is. Again, we get this crazy small number, but we log it so we arrive at something that is more decently suited for sorting.

If we do that for the covarying, for the *into*-causative as a whole, then in this case we'll get a different type of output, because now we have two slots. So we get everything sorted by collexeme strength, but then we get these word pairs, so in the corpus data we looked at here, *talk into letting, talk into surrounding, talk into staying*, those are the most strongly attracted pairs of verbs in the two slots of the construction. Then we have nice examples like *torture into confessing, shock into understanding*, whatever, all sorts of things like that. And then one can look are there particular cause-effect patterns in there, and for example, do those vary across different cultures, where of course we're not pretending this is a deep anthropological analysis or something like that, but it was a fun to look at the time. So the general semantics of the *into*-causative would be something like this, namely that the agent, the subject of the first verb, forces or tricks the patient, the direct object, into doing an activity that the patient would normally not want to do. If I blackmail you into accepting something, then normally you wouldn't have accept it, just like I have the power to blackmail you into that against your will; same with *force* or *bully* or *shame* or *embarrass* or these types of things, and we'll revisit the construction and its semantics at a later point. So some general characteristics of the pairs here of the verbs: there are forcing/tricking verbs in the cause slot, and there are sort of activity verbs in the result/effect slot, and to some extent at least like I said before, they instantiate some culturally-specific frames of entrenched cause-effect relationships. For instance, we often find things that belong to the commercial-transaction frame, to the confession frame, so to *terrorize into confessing* were the examples we found. And if you test American English vs. British English in terms of what happens in the cause slot and in the effect slot, then for instance we find things like this, in American English the cause verbs are, I mean a lot of times they are communication and they're physical force, and the patient often is restricted from doing something. So the forcing/tricking leads to the patient being less active. If you look at the same construction in British English data, you'll find quite different patterns, so in British English the verbs that are used here are stimulation verbs or negative emotion verbs, so verbs like *embarrass* or *shame*, so verbs that put you in a negative emotion so that you might either not do something or that you do something to get out of that emotion. And then we have things like threatening or physical force, so physical force is attested in both but it's much less prominent here. And then the other thing that changes is that patient is set into motion. So in American English the patient is restricted; in British English it's more like that the patient then does something to get out of that state. If we look at the effects then, the results are not that fascinating, to be honest: In American English there are a lot of light verbs, so actually it's very difficult to come up with any consistent pattern there. In British English, what happens a lot of times is the action that the patient is then beginning to do is a communication verb. But again I mean the effect size of this study is certainly not particularly great.

To sum up, last slide before maybe a short demo if it works here. The idea of collostructional analysis is to take an association measure from the domain of lexical association and apply it to the domain of lexical association either within a construction or to one or more constructions. For collexeme analysis, that means you look at words and constructions; for distinctive collexeme analysis, it means words to one or two constructions. The script that I wrote actually has an extension that can also be applied actually not just three, actually too even more, to one of however many constructions you want to look at. And covarying collexeme analysis then was to look at how words are attracted to other words within the same construction. Just like all corpus approaches, it kind of has to observe, is has to be based on observed frequencies and this case co-occurrence and quite usefully so, and of course rewards or emphasizes high co-occurrence frequencies but not unconditionally so, because those are normalized against high frequencies of things that are used everywhere. And the way this is done and this is going to be important a little bit later, that's why I try to make this point very precisely, we use the Fisher-Yates exact test for this measure but actually that's not necessary, you can use any association measures that you want, but we have good reasons for using that test and I'll return to that question at some point but theoretically—we have put this in writing, again this will be important later—it wouldn't have to be in that test. We have some experimental support for this from a sentence-completion task, or from a self-paced reading task and I will mention other things later. So we think it's really a nice approach, and the question that will lead to the next talk will then be "how can you not like this?" Unfortunately I will have an answer for how someone did not like this, which I will hope then to undermine, but that is a general idea. So this is the part of the talk, if possible, like I said I would like to show you this script and action for at least one application, so let me see whether I get this done.

Let me show this first, please. This is R, I've loaded the script which is called *coll.analysis*. When you press ENTER then, basically you'll get a whole lot of stuff like, where have the analyses like this being published? Then you'll get the usual "no warranty" and all that stuff, blah. And then it says "you should have received this program with collection of examples". I'm gonna run one or two examples, and so that you can see that, and then it says "Ok and press ENTER" to continue, so you press ENTER to continue. Then it says "well if you use it please quote it like this", press ENTER to continue. Then it says "Ok and what kind of analysis do you want to do"—let me have a look, which one might be best to do . . . Let's do a simple case, let's do a distinctive collexeme analysis and see whether that works. So you pick No.2, and then it says "Ok and the distinctive collexeme analysis blah", then it says what an analysis does, and then asks you whether you want to compare two constructions to each other like we did in the talk, ditransitive vs. *to*-dative, or *will* vs. *going to*-future or something like that, or whether you have in fact three or more alternatives like *will*-future vs. *going to*-future vs. *shall* or something like that, so in this case we'll keep it simple: we'll do two alternatives, so we choose the first option somewhat intuitive. Then it says "Ok and how many decimals do you want the output", and to keep things simple, let's just say 4, whatever. Then it asks you "which statistical measure you want to use", and the default would be the first one which means if you don't have good reasons—there is one good reason why you might not want to use the default, in this case we don't have that reason, trust me—so you want the default. Then, "how do you want the output sorted", again there is no good reasons to not take collostructional strength, because that's why you do the whole thing, so four. Then you'll get all this stuff, there are two different types of input format the program accepts, either it looks like this or it looks like that.

So this is the readme file that you get with this program, input format. I am supposed to enter one, so the input format is this. Basically, it's just a table that says what the word of construction is and then what word occurs with that construction. So raw list of all tokens and choose the input file, and that is 2a, right? So press ENTER to continue, and that opens this little thing so you choose the input file somewhere on your computer. Then it asks you for the frequency of the first construction, and note it gets the name of the construction out of that file, so you don't have to enter that. That frequency is something, 958, the frequency of the other construction was 814, so you just answer these questions. And then it asks you where you want the output, and for now I'll just put it in the terminal, so on the screen. So you say ok 2, and that's it. So then you get this basic output, and again another disclaimer about warranty and quality and everything, then there is the legend that says what all these columns are. Then we see, for instance, this is then ranked according to the construction to which a word is attracted and then by strength. So you can see here the preference of occurrence, first come all the ditransitives and then all the prepositional datives, and then within each construction it is sorted by collocational strength, so first all the ditransitives … and then here the number becomes big again that's because now other construction comes so we can see *give* and *tell* and *cost* and *show* are most strongly attractive to the ditransitives, and *bring* and *take* and *play* and *pass*—again that's from sports coverage, to *pass the ball to the defender* whatever, I have no idea—and so then you can interpret those results and put that in an Excel file and do whatever, and there is a small reminder down here that tells you what these values here mean, when those values indicate significance or not. So once you have the data out of the corpus, running that whole stuff is actually really really simple. Thanks.

LECTURE 6

# On Frequency in Corpora 1: Frequencies vs. Association Measures

So this talk is essentially a continuation of the one you've heard earlier today where I basically closed with a question "how can one possibly not like Collostructional analysis"? As I indicated this morning, there are actually people who failed to see the light and who dare to criticize this approach. So in this talk basically I want to recapitulate some of the points of critique that were brought up against the method I described this morning, and then basically produce or provide a rebuttal of many of these aspects, at the same time hoping to touch upon a variety things that have to do with how to look at frequency effects in corpus data. So this first part here will be mainly concerned with this issue of frequency vs. association measures. And then the first talk tomorrow will try to provide a more general picture of frequencies and things like that that we would want to look at that in corpus data.

So just to recap and especially maybe for those who were not here this morning: Collostructional analysis, which I will abbreviate as CA in the remainder of the slide, because it will show up quite a number of times, what it does it applies the logic of association measures—and again I will use this *AM* abbreviation from now on—it applies the logic of association measures to lexico-syntactic co-occurrence. So the idea is that we quantify in a particular way that I'll recap in a second how much does a particular word or how much is particular word attracted to, or repelled by, a syntactic pattern or a syntactic construction, if one wants to use Construction Grammar terminology? Crucially, and this will be very important later, the method basically involves four different steps. The first one is to retrieve all instances of a construction, a construction *C* if you will. And that might involve a lot of manual annotation. Again, if you have syntactically annotated corpus a lot of times you can rely on that, but even in that case very often manual disambiguation is necessary, but the idea would be to get basically all the instances of construction that are attested in the corpus or in the representative subpart of the corpus.

Second, you look at a particular slot in the construction that you are interested in and in most cases that has been so far the verb slot in a syntactic linking or argument structure construction, and you compute association measures for each of these collexemes, the word in that construction in this slot of construction. And then as you've seen especially in the demo this morning,

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_007

you rank these collexemes, so the words occurrence in the particular slot by their association score and then explore the top whatever twenty, thirty, I don't know, collexemes for functional patterns. And *functional*, I'm using it in a very wide sense so that includes semantic, pragmatic, discourse=functional and all sorts of patterns you might think are relevant for a particular construction. As I've indicated this morning, typically at least not necessarily but really typically at least, all these association measures involve 2×2 frequency table like this, where we have a construction in the row and then all the other constructions in the next row. So this means 'not'. We have a word in one column, the word of interest, and we have all other words, so 'not words' in this column, and then we have a 2×2 table where the overall sum of the table indicates usually the frequency or the overall number constructions in the corpus, and crucially as a colleague and I has said very specifically that any association measure can be used that you can apply to this type of 2×2 table. Most of the time, people have used the Fisher-Yates exact test which probably has to do with the fact that we think it's the best measure but also it's one that is implemented in my script which a lot of people have been using over the past few years. The advantages of it are that it has no distributional assumptions. For instance, a variety of measures require, like a $z$-score or a $t$-score, essentially for a significance test, they require normality, so a bell-shape curve like this, which you can pretty much forget about it when you deal with corpus data, I mean, that just doesn't happen. The Fisher-Yates exact test doesn't need that. Secondly, Fisher-Yates exact test is better at handling the low-frequency data than many other tests since in most corpora, for instance, half of the words occur only once, low-frequency items or something we consequently have to bear in mind. And this test is pretty good at doing this.

Third, and this is also important although it is a matter of controversy. The Fisher-Yates exact test is a significance test, so it returns a $p$-value, or a significance test, that usually would be smaller by five percent, and as such it correlates or reflects two particular types of information. On the one hand, it reflects what is called *effect size*, so that will be the answer to this question how strongly is the word attracted to a construction, or how strongly is a word repelled by a construction? That would be effect size. Secondly, it reflects *sample size*, which means if you have more data to base your analysis on, the value will be higher. So what I mean by . . ., this is what I try to reflect here, basically. If you have a distribution that something occurs somewhere fourteen out of thirty-five times, then that's basically two out of five. This is also two out of five, so eight out of twenty, right, that's the same ratio. What the Fisher-Yates exact test will do is it will rank this one more highly, because the same ratio has been observed in the larger data set with a larger sample size, and that is intentional, I mean, it's obviously a design feature of the test, we think that it is a desirable feature of tests to look at collocations because obviously when you are interested in how strong the attraction of something to something else is, or how strongly the entrenchment of some co-occurrence data is, then high frequency should help.

Now the logic of this type of analysis in this case for distinctive collexeme analysis like I said it is always based on 2×2 tables, and this morning we looked at three different types of methods. So collexeme analysis was the one where you basically measured how strongly are particular words attracted to one slot in one construction, distinctive collexeme analysis compared words and their attractions to two or more functionally similar constructions. And a lot of the work that has been done here has involved alternation research. And this approach—again, I want to make it very clear and you will see later why—so the idea is this serves to rank-order collexemes, so the idea is not really so much to test whether any word is in a construction significantly more often than expected or not. So that doesn't matter so much, what the point here is is to be able to say this word is attracted most strongly—whether that's significant or not is really not that relevant. Second, it normalizes the frequency of occurrence in a construction because words that are very frequent everywhere will be downgraded in collocation strength and as I said we use this particular measure because it can handle all of these things at the same time. Also it can handle attracted and repelled collexemes, so we've seen a case this morning where a word occurs somewhere less often than expected by chance. Maybe because it's highly frequent but actually in this particular construction pretty rare. There is a lot of work on this approach basically because of its, I think, methodological simplicity especially with the script and we actually have a variety, a large number of cases where this has been applied quite fruitfully.

How can you possibly not like this? Let's ask someone how she couldn't. Joan Bybee in her recent monograph in 2010 has a very unfortunately titled section, 5.12, where she—I do not want to say, discusses collostructional analysis for reasons that will become apparent in a moment—but she mentions it. And she basically brings up four different 'problems'. The first problem that she sees is that the Fisher-Yates exact test is a significance test, so something that tries to distinguish does something happen by chance or not. And here are two quotes that are related to this issue at least, one is "lexemes do not occur in corpora by pure chance" and "the factors that make a lexeme high frequency in the corpus may be exactly those factors make it a central and defining member of the category". Second question that she raised in this connection is how is *d* calculated, where *d* refers to, let me just go back to this here, where *d* refers to this cell in the table. So how do you fill that number, so what are the cases that are not the construction and not the verb, how do you get that number, which is generally a legitimate question, but we'll get back to this in a moment. So that's her first set of problems. The second one is that she says "no cognitive mechanism is proposed that corresponds to the analysis". So it's a cognitive=linguistic approach using corpus data, but then we supposedly don't discuss the cognitive mechanism that does this. Third, she says "since no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it". To be continued later with, "since it works only with numbers and not with meaning'. I am happy to admit at this point that into a collostructional analysis, the type of the way is done, semantic considerations do not enter into it. The rest of that statement we will see. Fourth, she refers to previous study of hers together with Dave Eddington and then claims that collostructions do not distinguish low-frequency semantically related items—so words attracted to construction—from low-frequency items that are semantically unrelated. Basically saying in a way, if something is low frequency, then collostructions are unable to distinguish between what we maybe want to know, namely semantically related stuff, and what we maybe would find unimportant, namely semantically unrelated items. And showing . . ., I mean she 'discusses' that by referring to this previous data set.

So let's pick this part. First point of critique, the one that has to do with, this is a significant test and how is *d*-calculated. This is the bottom line: Basically she misses the point completely, I think. Why is that so? Remember what she said. She said "while lexemes do not occur in corpora by pure chance"—no one would ever say that they do because if they did then it wouldn't make sense to compute association matters; it wouldn't make sense to look up frequencies because things are by chance, what would you learn from that? Obviously, no one ever would say that is the case. More specifically with regard to this particular question the use of significant test is a statistical heuristic, it is one way of normalizing for overall frequency and so on. Like I said before, we explicitly said "any association measure could be used", so we didn't say you have to use a significance test. If you want to use mutual information, which you shouldn't, but if you wanted to, be free to do so, be my guest. And in fact there are studies out there that have used measures that are not based on significance tests. So this is just off basically, because again what this approach, any association measure in fact, does is just downgrade words that are really promiscuous that show us all the time everywhere. It upgrades words that are highly faithful to the particular construction under investigation and basically in a way what it does is, it adds a dimension of information that frequencies alone can not reflect very well. So if you look at this graph here, forget the *y* axis for the moment, just look at this dimension, forget that these values go up. If you look at the frequency of a word in a particular construction, then obviously that can range from zero to whatever. If you're interested in ranking words according to the importance for a particular construction, then you will basically use the position of a word on this dimension. So the *n*, let's assume that's the verb, that would score twenty, this would score twenty five, this would score a little bit more, so for each of these words represented here by these individual letters, you'd only take the information on the *x*-axis into consideration, because you don't do association measure, you just look at the frequency. Now, that misses some information, namely, it misses maybe the information that a verb let's say like *c* or *b* is highly frequent everywhere, or more likely actually, that verb like *m* or *k* is highly frequent everywhere. So by adding now the additional information that comes from an association measure, so something that penalizes verbs for being highly promiscuous, you add the second dimension. So the information of the word *a* is not just the position of *a* on this scale but also on that scale. So in a sense quite trivially, geometrically you add a second dimension of information. So obviously in some sense at least, it kind of *has* to be more precise. What does that amount to in the particular cases we've seen? In one case this morning we have seen that we looked at the *as*-predicative, so this construction *will I regard myself as an actor* or something like was one of the examples. And what the analysis did and they use association measure, it made *regard* not *see* or *describe* or *know* most representative of the *as*-predicative. And why is that good, as we'll see later as this is actually good because *regard* semantically is very close to what this construction means and verbs like *see* or *describe* or *know*, they can occur in a huge number of contexts and a huge number of constructions—*regard* cannot, and it is exactly that affects that is brought out better by the association measure than by frequency alone.

Secondly, if you use this approach to compare to two different constructions, you also get something out of it. Namely, if a word occurs equally frequent in two constructions, then if you look at frequency only you'd have to say that verb or word in general is equally frequent, equally relevant to both constructions, but that of course would kind of presumes that two constructions are equally frequent, which most the time is not the case. So for example, if we compare *to* and *-ing* complement constructions, so *he considered stopping smoking* or *he considered to stop smoking*, then *consider* is equally frequent in these two constructions, but the *to* patterns is much more frequent. So the fact that the verb *consider* manages to squeeze the same number of itself into the rarer *-ing* complement constructions, says that is what it actually prefers, something that regular mere observed frequent is not able to show.

Now what about the Fisher-Yates exact test is more specifically? As I've said before, it has a variety of nice statistical characteristics in the sense that it doesn't require particular distributional properties—it is a significant test, yes, but unlike some others it incorporates frequency in such a way that entrenchment-based accounts what actually benefit from this, because as I've mentioned before, with increasing frequency of co-occurrence, collexeme strength goes up, so remember this fourteen by thirty-five compared to eight-to-twenty examples, where the higher co-occurrence frequency is rewarded by this measure, which is actually what Joan would want. Secondly, we find that as the frequency of something in a construction increases and as the frequency of a construction itself increases, we have sort of an exponential increase such as the higher of frequency becomes, the more higher the association measure becomes, and this type of non-linear relationship is something that you very often find in learning curves and forgetting curves and stuff like that, I will come back to that a little bit later. Also, Fisher-Yates exact test is a relatively complicated statistic but of course you can't really use that against the approach, because, you can't say it better than Anatol did, "statistics don't have to be simple to provide accurate representations": if the underlying process of what is modeled is complicated, then maybe a statistical measure has to be complicated as well. In Daniel Wiechmann's comparison of a lot of different association measures with the psycholinguistic reference data, Fisher-Yates exact test actually came out second best and the only measure that is better is one that called *minimum sensitivity*, which is a theoretically highly problematic for reasons that I'm happy to elaborate on in the Q&A but I don't want to clutter at the talk at this point.

Now don't look at the slides in front of you, just look here. So now about comparing Fisher-Yates to some other measures, I am coming back to a question that was asked this morning. So here are three different rankings of ditransitive verbs according to different measures. One, so this is one ranking based on this measure, a second ranking based on these results, and the third measure, third ranking of verbs based on this measure. Now if you think of the ditransitive and what do you know about the semantics and stuff like that, which of these rankings do you prefer? There is only one possible answer. Obviously, the first one. Someone this morning asked about mutual information, so this is included here, but it is the third one, and this is one you would probably not want. So why is the first one, Fisher-Yates, the best, I think? If we start from the right, then here mutual information basically shows exactly the type of problematic characteristics that we've talked about now a few times already, namely it accords really high, prominent to things that are actually really infrequent. Right, I mean, *accord, award*. then there is *give* but then *allocate, profit*, many of these verbs are certainly not particularly frequent in everyday discourse. So a measure that takes those and says that these are the verbs that are most representative of the ditransitive, probably has statistical characteristics that you don't want to rely on that much.

Now log odds are somewhat better in the sense that at least it gets that *give* should be the prototype, right, which is in every analysis, the ditransitive would decline, but then look at the small distance to the next verb. And again, the next verbs are ones that are relatively infrequent, maybe not exactly core ditransitive verbs. Now if you look at Fisher-Yates exact test, it's a completely different story: As you would probably want to *give* wins by a very wide margin as the absolutely most prototypical verb, and then we have *tell, send, offer, show*, so high-frequency verbs with a transfer-related type of semantics that are extremely compatible with previous analysis of ditransitive. So association measures of this type should actually be highly appealing to cognitive linguists, they result in frequency and their effects relations that are non-linear as I have shown you in this plot on the right hand at some point, just like frequency effects that are non-linear, just like learning curves that are non-linear, like forgetting curves, like priming decay studies, all of these are non-linear and, counter to frequency, this is what association measures return as well. They reflect things we have known from psycholinguistic studies and also from associative learning theory studies, namely that raw frequency of occurrence is less important than the contingency, the conditional predictiveness between a cue, which could be a construction and an interpretation, which could be a verb in this particular setting. Nick Ellis and a colleague in the *Annual Review of Cognitive Linguistics* put it very nicely, "contingency and its associated aspects of predictive value, information gain, and statistical association, have been at the core of learning theory and then also later linguistics ever since".

What about the final point that Joan Bybee brought up? So this question of how is the *d* cell computed, so the cell where you have 'not the verb in question' and 'not the construction in question', I mean it is computed for every association measure namely on the basis of approximation that make sense for a particular corpus. So obviously if you look at let's say transitive constructions, then the cell *d* will not be computed as a function of how many prepositions are there in the corpus because the two things are not related. Obviously you would choose something that is related like fine-lined verbs or verbs frequency something that approximates frequencies of constructions at a similar level of granularity. However, there is also good news namely that simulations show that collostructional analysis rankings are extremely robust even if you have no clue whatsoever which frequency to take. So here is an example. I took ditransitive data and this first row and this first column, those reflect the data that were actually found. So it is a corpus size of a 140,000 constructions and observed frequencies as they were found, and then I just increased corpus size by a factor of ten, so basically making up the number that would go in there, and here I take the real corpus size but first take half of all the observed frequencies. So basically subjecting the real data to really severe changes, and here I do both in this part of the table. As you can see, however, the resulting rankings are extremely highly correlated with each other: manipulating the data this way, so the uncertainty that comes to estimating how many verbs and construction are in there actually plays much no role whatsoever. The rankings will turn out to be the same pretty much the same all of the time.

So what then if we pit frequency on the one hand against association strength as operationalized here on the other hand? You've seen this type of results before this morning and also earlier in this slide show: *regard* is returned as the central verb although, for the *as*-predicative although verbs like *see* and *know* are as frequent or even more frequent, but they are also more general so they have a less degree of discriminatory power for this construction. We looked at the result from the sentence completion task this morning and the self-paced reading task. Nick Ellis and Rita Simpson-Vlach, they looked at association measures compared to frequency as a predictor of how speakers rate the formulaicity of constructions and the association measure did better. Here is the self-paced reading time result again that you've seen this morning. Timothy Colleman and Sarah Bernolet did a study where they looked at the verb-specific preferences in the Dutch dative alternation and they actually put that very nicely: when you look at the results and in terms of frequency, then it makes no sense whatsoever. It seems completely arbitrary. Once you look at the same data through the prism basically of an association measure, then everything falls out very neatly and exactly the types of verbs you expect there show up at the top.

All this is not supposed to say that we couldn't do better. So Fisher-Yates exact test is a measure that has yielded very good results but it may not be the best measure that there is. I want to at least briefly mention today and later in another talk a measure that might be better, and one problem about this measure of Fisher-Yates is that it's bidirectional. So it quantifies how strongly a verb and a construction are attracted to each other in both directions. It is not sensitive to maybe the verb likes the construction more than the construction likes the verb. It cannot handle that. And that of course a way is cognitively not particularly realistic, because the way we learn things is probably relatively directional, we go from something we know to some new thing, so there is an inherent directional bias. So a unidirectional association measure might be more useful and here is the one that one might use, it's called ΔP, and so it can take both directions namely ΔP of one thing given the other or ΔP of the other given the one. And this is how it is computed, again I would not discuss in great detail right now. I just want to show you one example, namely, the collocation *of course*, which is so strong a collocation that the people who compiled the British National Corpus actually made that a multiword unit. So it's not just tagged as 'there is a word *of* and there is a word *course*' and it get its own special tag, actually you should treat that as one word, which kind of makes sense, right? Now if you apply all sorts of normal association measures to this then you get these results, all of which show that this is a super strong collocation. But what all of these measures miss is that that is actually only true in one direction. This side of this line here indicates if you have *of*, there is a ton of things that can happen after *of*, it's probably the second or third most frequent word in English, I mean *the* is always the first, then we get *of* or *in* or something like that. So after *of*, a lot of things can happen, so the association is really low. But if you see *course* in a corpus, there is a really high chance that the word before this will be *of*, there is not a lot of things that happen before *course*, so this measure can bring this out nicely, and it will be cognitive more realistic, because it incorporates the directionality. As you will see later this week, this discrepancies like this, where association is really only stronger in one direction, they are all over the place and I looked at the sentence completion task data at some point later using ΔP as an association measure and it is in fact a very significant predictor of what the subjects did. So ultimately down the road this might be a very useful way to basically improve even on Fisher-Yates.

Second point. So Joan said 'no cognitive mechanism has been discussed.' That's just not true. If she had read the papers, she would have seen that we do. So on page 237 of a paper that she discusses, we talked about the psycholinguistic studies of language acquisition quoting earlier work by Goldberg in this connection. We discuss the relation of collostruction strength to raw frequency in that connection. We make connections to the notion of cue validity, which is a very prominent in prime in the Competition Model, referring Goldberg's and Ellis's work. We discuss, two pages later, we discuss the relationship of collostruction to entrenchment. So all of the stuff is actually completely in there. In Daniel Wiechmann's study, again, he compares association measures also to cue validity and to cue strength and Brian MacWhinney said, "collostructions is exactly what we need for this." So the fact that we don't discuss any cognitive mechanism might be involved here is simply not true.

Number three. The idea that 'no semantics enter into it' because... 'no semantics enter into the analysis and, thus, no semantics can emerge from it'. And it's actually curious that Joan would say that, because on other occasions, she would pretty much say the opposite. If you look at computational psycholinguistics, or just psycholinguistics in general, or I sometimes call it distributional linguistics, there is a lot of work that has shown that if you have a statistical algorithm work only on completely numeric totally semantics-free input, what you get out of it are functionally, i.e. semantically, highly coherent clusters. I only want to mention two studies here that have shown this to a very large audience. So Reddington, Chater, and Finch, they looked at co-occurrence frequencies of 150 bi-grams before 1000 target words. So I mean no semantics whatsoever entered into this analysis, just frequencies of how often does this happen before that. And then they did a cluster analysis on this, which is a type of analysis if you remember that returns these tree diagrams that show how things relate to each other, and the cluster analysis, what it returns essentially were parts of speech. And in a cognitive grammar approach, parts of speech of course have semantic or functional import so this is a clear case where no semantics went into it, but something that is a functionally very relevant came out of it. Second case of very similar study to the one by Reddington, Chater and Finch, just use different definition of contexts. And again, it found highly functionally loaded output based on something where only numbers went into this. And computational linguistics per se, so not psycholinguistics, there's also a lot of work that has shown that you can do a lot of things by just entering co-occurrence frequency into a system. So latent semantic analysis for instance would be one point where you can clearly see that the co-occurrence information in large texts, helps, for instance, a lot in finding out what the text is about, like for information retrieval, document summarization and things like that.

The second point that I would like to make in this connection is that in collostructional analysis, the semantic analysis *follows* the statistics. So semantics don't enter into doing it, but they certainly come out of it. And actually there were quite a few studies at the time that Joan wrote this that show this. So we've looked at the ditransitive example earlier, so the verb that was returned as prototypical/central is *give*, and that was discussed to be very similar to the semantic analysis of Adele Goldberg, so that's already semantics. Then as I've mentioned this morning, the next verbs that follow in the collexeme analysis are all the senses that Goldberg's analysis of ditransitive has posited. You can basically just read them off of the ranking of the verbs and if that is not semantics, what is? And then in a paper that was published only 2010, but that was publicly available since 2006, I think, or 2005, we did something that even better than that, namely, we did cluster analyses on co-varying collexemes. So remember co-varying collexemes was these types of analysis where you look at one construction, but it has two open slots, like *to VERB someone* into *VERBing*. OK, so what we did, we clustered the verbs—the *to VERB someone*—and clustered them on the frequencies of the *into VERBing* something. So remember this morning we have seen examples of *blackmail into accepting*, or *force into marrying* or something like that. So we would cluster the first verbs, *blackmail* and *force*, on the basis of how frequently do they occur with the other things later. And we did the same thing with the *way*-construction. We clustered the verbs in the *way*-construction, *to make/fight/wave your way*, on the basis of the preposition of the paths. So you *make your way to the top, you fight your way through the crowd*, those we used. If you do that, we get something like this. So the verbs here are these *forcing* verbs and *tricking* verbs, stuff like that, they are cluster on the basis of what the patient then did. And as you can see there's quite a lot of cluster structure here and so there are some black boxes here and they come in quite revealing clusters. So here we have a cluster that contains physical force verbs, I mean *pressure, pressurize, force, push, bounds, press*. They all have in common that that's potentially or actually rather physical activity. Then here we have a cluster with the exception of *talk* very nice and clean cluster of trickery verbs and then we have two clusters of stimuli verbs, namely positive stimulus verbs and negative stimulus verbs. These are positive stimulus verbs in the sense of that they provide positive incentive to do something, and these are ones that provide negative incentive to do something. And all this semantic structure just come out of co-currency frequency. So, no semantics goes into it, but a lot of semantics comes out of it. We get a very similar nice result, although different in structure, from the *way*-construction, so we have two highly frequent all-purpose verbs that occur so much in everything and then again we have a bunch of physical force verbs, and then we have three different types of slow-motion clusters. Here we have things like *claw, grope*, so stepwise things where you move along a path like this, then we have things like *wind* and *thread*, *weave*, so more curvature stuff like that, and all of that is just on the basis of co-currency frequency with preposition so no semantics went into it, but semantic patterns definitely come out of it.

Number 4, the perceived lack of discriminative power. To remind you, Bybee said that low-frequency collexemes do not distinguish between semantically related and semantically unrelated items. And she refers to this study on the Spanish verb—it's not like that I speak Spanish, *quedarse* and *ponerse* or however you pronounce that. So how did she show that? That's actually really interesting, I was amazed. Remember, these are the steps of a collostructional analysis, right? So you retrieve all instances of construction; you compute the association measures for all the collexemes that you find in there; and then you rank-order all these by the association measure, whatever it is; and then you look at the top 10, 20, 30, I don't know. The 'analysis' that was used to make this claim was a little different. They took 24 adjectives that occurred with these verbs and computed association measures only for those, and the frequencies of these verbs were one or less. So, all the high-frequency things that usually in a collexeme analysis get pushed to the top, because they occur frequently with something else, they didn't even enter into the analysis, so she didn't take all instances, she didn't allow for high-frequency verbs, uh, adjectives to even make the cut because those were not included—in fact, she focused on hapaxes and things don't even occur with what's she interested in. If you basically do like half a step of the four-step analysis, you don't get a result with discriminative power, I grant you that. But basically what she did—she ignored all previously published results that did show that there's some discriminative power and she ignored how the analysis she is criticizing is actually done. The semantic analysis is done *after* the stats, namely, when you have the ranking, not *before*, you include *all* collexemes, not just those that occur maximally one time and she even does not discuss, which is a minor technicality here but still, she only correlates the results that she gets with positive acceptability judgments, not with negative acceptability judgments, although there is work that has shown that collostructions can predict those, too.

Now if you do this type of analysis the way it was actually meant to be, what do you get? So you end up expecting words that ranked highest based on something that is composed of both frequency information and contingency, remember Nick Ellis' quote. So in these two studies, we look at ditransitives, dative alternation, *as*-predicatives, and I've shown you these results above. We have applied this to *into*-causatives and again you see some of the results and the results we got were better than a pure frequency-based analysis by Hunston & Francis. Again Nick Ellis has done extremely interesting work—I recommend him for the next International Cognitive Linguistics Forum here—he has done very interesting work in the domain of second language acquisition. For instance, he has shown the degrees or the frequency with which learners get a new construction and use it is predicted by frequency, ΔP, and $-\log_{10} p_{\mathrm{FYE}}$—but actually Fisher-Yates exact test outperforms frequency in 2 out of 3 constructions that they study. I've shown you this one before, so sometimes it's only the association measure than actually makes sense of an otherwise frequency-based ranking. In much earlier work, Gregory et al. (Michelle Gregory, Dan Jurafsky, Alan Bell, and some others) have shown that association measures predict pronunciation reduction effects, whereas frequency does that to a lesser degree.

There are some additional difficulties and now I'm beginning to prep the next talk in a way. So one thing that Joan said is that collostructions ignore low-frequency collexemes, which may reflect productivity. In a sense that's a half-way valid point. So a collostructional analysis the way that I've introduced here will place a larger degree of emphasis on high frequency items, because those will be ranked more highly in the statistics. And it's true that low frequency items reflect productivity, many of the productivity measures in the morphological productivity literature by Harald Baayen and others, Jen Hay, Antoinette Renouf, they use hapaxes actually to quantify productivity. However, first, this is not really the goal of collostructional analysis: it has never been claimed to be an analysis that measures productivity. So that's kinda like saying "your experiment doesn't reflect corpus frequency". Well, yea, it's a different type of tool. Secondly, if you do a collexeme analysis, of course, you get the result, I mean, you've seen this morning; you get the result for all the verbs so, if you want to use the results for productivity, you can—it's just not something that has been prioritized. Second, she says, 'it's not permanent how often a lexeme does not occur in a construction'. So, basically she says we need to know how often *give* is in the ditransitive, but it doesn't really matter how often it occurs elsewhere. But as I said before, we've seen that the literature on learning says otherwise, because it focuses on contingency or predictive value or predictive power of something, so obviously, if some observed frequency of *give* in the ditransitive is really high and it hardly shows up anywhere else, then that is a very predictive pattern, whereas if *give* shows up there a lot, let's say, *see* shows up in the *as*-predicative in particular number of times, but is super frequent everywhere else, then that is not predictive; obviously that is interesting information. And actually, collostructional analysis uses information that Bybee on other locations in her book actually uses herself. So her example (9), I think, from chapter 5 or 4, where she talks about that *drive someone mad/crazy/silly* or something like that idiom. She doesn't look at how often *drive* occurs somewhere else, but she does look at the type frequency, sort of various different alternative things that something might be used with in these different slots. Essentially, the collostructional analysis basically does that, it just systematizes it by normalizing any of these expressions against their overall frequencies. So this example is a simplification, but her example is instructive. However, what I think points to is something other than what she concludes, namely the situation, the type of data that we deal with are actually *way more* complicated than both Bybee and collostructions assume. So collostructions, also as I will argue in more detail tomorrow, is also just a very crude simplification. But simplify is a little bit less than you just look at frequencies. What we really need is something much much more complicated and that's what I'm going to talk about tomorrow. If you look at a particular construction, what we actually need is the type frequencies and token frequencies in all the slots of the construction. So if that construction has just one relevant slots, so like ditransitive for the verb, then we don't need to know that one verb occurs there this many times. What we need to know is, how many verbs occur in that slot, the type frequency? And for each of those types, we need to know how frequent is it in that slot? Second, we need to know about the dispersion of the tokens. So if something occurs in a particular slot of a construction, does that happen everywhere? I mean, is that a productive pattern that every speaker would use and recognize and produce, whatever, or is something that is very specific? I just had a fun discussion in Stanford the other day what we talked about leaving out otherwise obligatory direct objects. So one case in point for this would actually be the intransitive use of *lift*. So usually if we use verb *lift*, we have a direct object, *lift the microphone, the computer*, whatever. But there's a frequent intransitive use of *lift*, that is, however, highly under-dispersed: it's restricted to particular type of community, namely the community of body building where it is clear what they *lift* because they are part of that community. So you find online chat rooms—that student was looking at stuff like that—she found people in online chat, they would type, "OK, I gotta go, I'm gonna go lifting", leaving out what they lift. But that's a body building on-line community—what are they going to *lift*? Weights obviously. But that doesn't mean the intransitive of *lift* is highly dispersed and that everybody would do that, it's a part, to some extent at least, a function of which community you are a part of. Will everybody in the community be likely to recognize what you are saying? So we need to know that to what degree are the token frequencies that we find actually distributed evenly throughout the speech community. And then we need to know about the entropy or the distribution of the token frequencies, and most of the time those will be Zipfian. That means a few verbs, or words, will occur, will account for a high share of what happens in a construction and then the rest will taper off and a lot of cases will be really rare, And that of course has implications as I'll discuss tomorrow for learnability of the constructions and for their productivity. And then we need to know of course the exact frequencies and association strengths of elements to the slot they occur in and maybe what happens in other slots. And as if all that wasn't complex enough: if you want to do it really right, it would actually have to be sense-specific. So a lot of verbs—I am using verbs all the time, because that's what most collostructional analysis has been on that, of course, it could always be words -= most verbs have more than one sense. Remember the example of *run*, it has 50, something whatever. So computing the association of, let's say, an intransitive use of *run* to a particular construction is pretty pointless, if you don't include the information which sense of *run* is that actually, because some sense like the prototype will be really frequent, others will be highly infrequent. If you don't do your analysis in the way that accounts for these different polysmous senses, you are kind of just hoping that it will pan out in a long run, because you miss out the level of precision. Now in general, none of this Bybee-bashing should distract from the fact that on 95% of all issues we completely agree. We both think that usage-/exemplar-based approaches are the right thing. We both agree on the fact that frequency in general is an extremely important concept. We couldn't agree more on the assumption that much of linguistic knowledge and use and processing is affected by domain general mechanisms. I have mentioned some of those in previous talks, like analogy/similarity, when we talked about phonology of idioms, chunking, frequencies of exposure and everything. But as usual the devil lies in the detail and we have to be extremely careful to not throw out a method or its implications with the bathwater. Basically what we shouldn't do is avoid the true complexity that our data actually have to offer. For linguists, that's kind of bad news because that does get much messier and more chaotic, and more unstructured a lot of times than linguistic data. Tomorrow I will hopefully be able to elaborate a little more on this. Thank you.

LECTURE 7

# On Frequency in Corpora 2: The Broader Picture (Dispersion, Entropies, Zipf, . . .)

So, today's talk will basically be a continuation of the last two. So the good news is it is a continuation. The bad news is that basically now fun is over. Now we are getting serious.

We'll pick up the pace considerably today. In terms of the types of the statistical approaches we will consider. Basically, what I want to do in this talk is to demonstrate or to exemplify a cline of complexity of frequency information that I think in corpora should be considered more often than they are. So essentially, we'll start from something that we've already seen, but then move on to more complex types of co-occurrence information and what they can reveal. And the idea will be to basically say, ok, this is what is usually done, but then we should be doing this but even if we do this, we need more types of information and so on—so build up and elaborate on a variety of things that previous talks have touched on only briefly.

So basically we'll look at six or seven different types of co-occurrence information that I think will be useful. And the simplest one of them is that we look at the frequency of something in a particular corpus, or in a particular corpus part, so that would be the simplest type of corpus-linguistic information that you can get: raw counts or frequencies of a particular phenomenon, let's call it A, I mean that could be a word, that could be a construction, whatever. And that is a test that an X which could be a corpus or a corpus part, or a register or something. And the simplest way to get that information would be in a frequency list and schematically, with a particular example that would look like this, we'll have the frequency of something in a corpus. So for example, that could be the frequency of the verb *give* in a corpus and that could be 112, something like that. In the course of this talk, I will successively build up this type of frequency table, co-occurrence table. You should pay attention to how this thing grows as we look at more complex information.

Now it's obviously very crude tool, I mean it doesn't get any simpler than that, in terms of what you can get out of a corpus. But they're still quite important in a sense because they correlate with a variety of things, some of which are cognitive linguistics in nature and some of which are not. So entrenchment as discussed by Schmid (2000), for example, would be one particular case where corpus frequencies are cognitive-linguistically important. We've seen

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_008

how frequency of words can help, or can lead to, phonetic reduction or the development of new forms over time. We've also very briefly seen cases where high frequency of occurrence in the language makes irregular forms resistance to language change and so on. Obviously, things that are more frequent are probably easier acquired and acquired earlier in first language acquisition. Frequencies correlate positively with reaction times, actually negatively, sorry, the more frequent something is, the smaller the reaction time. So, lexical decision tasks and so on. So there is really a lot of different ways in which token frequencies as, however, simple as they are, can be quite useful.

Now, one problem of frequencies is the issue of dispersion that I've mentioned before in another talk. And that is because frequency per se is not a particularly, I mean, it's only one dimension of frequency information that we represent mentally. In corpus linguistics there was one study that met that point. The quite clear using in an example, the example of the words *HIV, keeper*, and *lively*, and if you look at how frequent those are in corpora, and in a particular corpus in the *British National Corpus*, 100 million words supposedly representing British English of 1990s. Then they all occur approximately 16 times per million words. So their overall token frequency is pretty much the same for all intents and purposes. However, if you divide the *British National Corpus* into one hundred equally sized parts, so one hundred parts in a million words each, then the word *HIV* occurs only in 62 of those while *keeper* and *lively* occur in 97. So obviously this word somehow is more specialized, more restricted to occur in only particular parts of the corpus.

There are dispersion measures out there that attempts to quantify this type of distribution, and if you compute one that has been frequently used, Juilland's *D*, then that value for *HIV* is much smaller than the one for *keeper* and for *lively*, indicating that these two words are more evenly distributed in the corpus compared to *HIV*.

There's a ton of measures that have been proposed. These are some of the ones that are maybe used a little bit more frequently, but a lot of them come with a variety of problems. One is that some of these measures require that the corpus that you have is split up into equally size parts. And a lot of times that is not really realistic given the way corpora are compiled or have been sampled. Secondly, somewhat interestingly, some of these depend on the order of the corpus parts. So if you take the BNC and you take, you follow the file names in an alphabetic order then you get a different measure than if you do it the other way round, which sounds like really something has gone seriously wrong, but there is a reason why sometimes that might be interesting. Second, uh third, some of these measures are really very very sensitive to zeros and outliers so a few corpus files that do not contain the particular item under investigation might already throw that measure off. Or they might be too insensitive, which means they return maximum values really quickly. So even that something is still quite under-dispersed that not distributed evenly a measure might already say: "oh, yeah, that shows up everywhere." So we basically need to find a measure that is sensitive enough but not too sensitive. And some of these measures have incomparable ranges, so some of those range from zero to one, some of those range from zero to infinity, so it's quite difficult to compare those different results.

Now, a measure that I have kind of developed and that I think handles all of these issues is what I called *DP*, for deviation of proportions. And one very nice thing about it is that compared to most of other measures out there, it's extremely simple to compute. Basically, you take three steps, namely, you compute—if you have a corpus and it has different parts, and first you compute the corpus at the size of each corpus part in percent. So, if you had a corpus that consists of ten equally sized parts, then you would have just ten times ten percent. OK?

Second, you compute the frequencies of the word that you are interested in in each part, in each, in percent. So if you have five parts of a corpus, and half of all the occurrences of this word is in the first component, then that would be fifty percent. And then you just compute the absolute differences between these two series of values. So *absolute* meaning you take away the minus something like that. You sum them up and divide them by two, then you get a value that ranges from approximately zero to approximately one, and that is, the value is oriented such that if it is high, then that means the word is not distributed evenly. Whereas if it's low, if the value is close to zero, then that means the word shows up everywhere with pretty much the same frequency.

Now, why would I want to do this? Well, here's one reason why. So here's a bunch of 68 or something selected words. We have the frequency of the word on the *x*-axis, and we have the dispersion on the *y*-axis. You can see that, of course, on the whole there is a relation like this, right? I mean, something that is highly infrequent cannot be distributed evenly throughout the corpus. So obviously, if things are infrequent then they are under-dispersed, and if things are very very frequent like, look at those function words here, of course, the words like *the* and *to* shows up everywhere. But the crucial point is that here in the middle range, words can have pretty much the exact same frequency, but they can be totally different in terms of their dispersion. So something like *hardly* or *anywhere*, while those are things that show up in a lot of different contexts, in this corpus, the word *Egypt* actually has pretty much the same frequency as the word *hardly*, but, of course, that is much more restricted.

So any study that only looks at frequency runs the risk of comparing apples and oranges, because words differ very much in terms of their dispersion.

Now, how do we apply dispersion? A lot of times it's computed over something that linguistically at least is irrelevant, namely files, right? File, corpus file, I mean, that is not a linguistically relevant notion, that's just a sort of sampling relevant notion if anything. So what we usually want to make use of is the fact the corpora usually have a linguistically meaningful substructure, namely something like registers, genres, or sub-registers, or things like that. So for instance, here, this is the hierarchical structure of the British Component of the International Corpus of English, which distinguishes spoken versus written as most corpora do. But then within speaking, we have a register distinction between dialogues, monologues and what they called mix. And in writing they have the distinction between printed and non-printed material. And then within each of the registers we have sub-registers. Spoken dialogue that might be private or that might be public. Spoken monologue might be scripted, prepared or unscripted. Written printed might be academic writing, creative writing, instructional writing and so on. So this is not files, but this is something that linguistically might very well lead to important differences. That has an important implication though, and that is if you ever write something up about how frequent something is in a corpus, then you are generalizing over all these different parts. If you say this word is this frequent, or this construction is this frequent in this corpus, then you are basically glossing over the distinction speaking versus writing. You're abstracting over register distinctions over sub-registers distinctions.

So, and then the thing is that that generalization you make about how frequent something is in a corpus, that may be valid, but it may also be completely off, because essentially you are pretending if you say:" OK, this word is this frequent in a corpus", then you are implying at least that making a distinction between speaking and writing isn't necessary. But of course it may well be, and you only know that if you actually look at it. And so your null hypothesis of "I only need to talk about this one level of resolution" may be terribly wrong. And I am going to show you an example that I think drives this point home very clearly, when you look at the frequencies of present perfects in a particular corpus. I've no idea why one would want to do that, but there was a paper that did that and that I thought made some interesting points so I replicated it here.

This is a graph that shows the frequency of present perfects in writing and speaking in that corpus, and again this is a box plot that sort of has the percentage of how many verbs were present perfects, so in writing the median was a little bit more than two percent, and speaking was a little bit more than three percent. The fact that these notches don't just overlap says that this is a significant difference. So if you're more like corpus linguistics, you might say "oh, this boot, I have significant difference way. I can write a paper". Writing is less than speaking, so I'm done.

However, if you take the written and the spoken data and you divide them further into registers according to the division of the corpus—now the white boxes, that is the written data, and the black ones, those are the spoken data—then you find actually there are significant differences between the registers. What looks like a relatively homogeneous block of writing here is actually, I mean, writing in itself has significant differences between printed and non-printed, because those don't overlap. And in speaking, dialogue is actually the same as monologue, with regard to frequencies of present perfects, but it's very different from what they call mix. So, second paper, significant differences.

But then there is some white space on this slide So I guess I have another point. You're looking at the same data, this time in terms of sub-registers. And again the white boxes here are printed as the written data and the black ones are the spoken data. And so we see that actually even something like printed or non-printed is not a homogeneous whole, there are significant differences between those two. And in this particular case, we can even see that actually the spoken data are quite homogeneous. I mean they are all on the same range with the exception of this one whereas the written data are all over the place from something as low as this, up to something as high as that, which is interesting in and off itself because most of the time people are like "oh, my God, spoken language is so diverse." In this case is what written language that is.

So what that means then is that frequencies in corpus data they should always be checked actually with regards to the homogeneity of the corpus, because any level of corpus granularity basically may give rise to very significant different results. Any of these might be worth of a paper. But any of these might be either compatible or incompatible in your hypothesis. But if you don't look at that, you'll never know. So basically the key here does not go with speaking or writing that is what everybody does, maybe also explore the final level of resolution that the corpus if it has that to offer.

Now, if you've seen however also that frequency per se are not always that relevant anyway. So, for instance, even something like reduction effects are maybe not so much due to overall frequency, but also to something like cumulative exposure on the one hand, but also contextual predictability on the other hand. More radical work by Harald Baayen not too long ago suggests that maybe frequency effects in general are really epiphenomenal, because frequency is correlated with a lot of other lexical characteristics, frequency of words. So he is suggesting—I'll not discuss much of the details here—but he suggests that actual contextual measures like syntactic families size (so how many different construction does something occur in?), syntactic entropy (how predictable is something syntactically?), overall dispersion, of the kind I showed earlier, all these types of things maybe much more relevant than frequency, I mean lexical frequency as measure than a corpus per se. So if that is the case then I guess adding context to a frequency counts would be a good idea. So why don't we do that? Get into the second level.

Now, we are not talking about the frequency of A, something in a corpus, but frequency of something in a particular context in a corpus. One way to look at this would of course be collocations, collostructions, colligations or something, where you look at a word in a context where the context might be another word, might be a pattern, it might be the position in a paragraph or in an intonation unit, anything like that. So that table now changes from just one frequency overall in the corpus, it changes to a frequency of something in a particular context. So it is not overall frequency of "give" any more, but "give" in a particular construction in a particular corpus.

And again, this has a lot of different applications, and again they correlate for instances with reduction phenomena, with grammaticalization. We say "I'm gonna" only with the future meaning, things like that. But we have seen already in the last two talks that if we look at the frequency of something in some context, it also always helps to consider the frequency of that thing in competing contexts. Like 'not in the ditransitive construction'. Actually I'm gonna talk about this much because the last two talk has already made the point, or tried to make the point, that that type of information would be more useful.

So we immediately go to number three, namely, that we should be looking at the frequency of something in one context as opposed to other contexts in that corpus. So we again add another level of resolution, moving on to the types of 2×2 tables and association measures that the last two talks talked about. So the table now becomes this. Maybe going back and highlight the contexts or the contrasts. So here we have "give" in a particular construction. Now the idea is: we look at "give" in one construction as opposed to the other and "not give" in one construction as opposed to the other, basically, taking the type of collostructional approach advertised in the last two talks. And that can be done using percentages or conditional probabilities. It can be done with the bi-directional association measures of the type that I talked about yesterday, like Fisher-Yates exact test or other things. It could be uni-directional association measures like ΔP which I will come back to in a moment. And again, if we use these types of measures, they can be useful in a variety of contexts. We have seen that they are indicative of the core senses of a construction, that they correlate with priming effects, that they correlate with acceptability ratings and sentence completions, all sorts of that different things.

Now what we have seen yesterday too was that the Fisher-Yates exact test as one association measure can get good results, but we maybe make it better than this. So the point there was, and I mentioned this briefly yesterday with one example, was that just like most other measures this type of approach is bidirectional but learning of things is usually not. So what I want to talk about today in more detail than yesterday is this notion that maybe uni-directional associative measures are more revealing.

This measure ΔP was first discussed in linguistics by Nick Ellis in a theoretical paper on the relation of linguistics and associative learning. And it's computed actually rather simply, it doesn't look like it, but it is. So ΔP is the probability of some outcome given that something else is there minus the probability of that same outcome, given that that cue is not there. And I'll show you how to do this in a table in a moment. That value again is again zero when the outcome of the cue are not related to each other. And that value is greater than or less than zero when the presence of the cue increases or decreases the probability of the outcome. Thus, in a way what it does is, it normalizes conditional probabilities, it's really easy to obtain because it basically just computing one thing minus another thing, and it has the charm of being made cognitively somewhat more realistic.

So what's it look like? If we have a table like this, which by now you know. So one word might be there or not, and another word might be there or not. I'm not gonna have this a, b, c, d corpus frequencies. Then this is the formula for ΔP in one direction. So it's the probability of the second word, if the first word is there, minus the probability of the second word, if the first word was not there. So it's basically just this fraction: $^{a}/_{a+b}$. So, this, divided by that, minus, this, divided by that. That's it. And ΔP in the other direction just transposes the table. So W1 given that W2 is there minus W1 given that W2is not there. So a divided by this, minus b divided by that. That's all.

So the example that you've seen yesterday, of course, if you do that on the spoken part of the British National Corpus, 10 million words of spoken British English, you get a table like this. So the corpus that spoken component has about 10 million words, *of course* is in there 5610 times *course* is in there this many times, *of* is in there that many times. So the probability of *course* given that there is *of* is just this divided by that minus this divided by that. And as I've indicated yesterday that is really really close to zero, because the word *of* doesn't really predict *course* very well. If you do it the other way round, then we divide this number by that one minus this by that one, and that value is pretty high, meaning that *course* is really a good predictor of *of*. And we've seen yesterday that this is much more precise than what usual bidirectional measures say, which they say "oh, yeah, sure, that's a strong collocation" but they don't reveal the direction.

Now this seems like a great thing, but we need to validate this idea. So there are two steps of validation that I use here. One is by looking at words where we would expect strong collocations. So I looked at 262 two-word units that corpus compilers have annotated as such in the corpus. So things that are spelled at as two words, like with the space between them, but that are so frequently used as one that the corpus compilers said "OK, we should mark that as one". *Of course*, this is one example *according to* would be another example *because of* would be one, *out of* in a lot of cases is one. So all these things, where you spell them as two words but actually use them as one all the time. And so for something like this, you would expect any collocation measure that is worth anything to say, there is some big dependency there. If you look at the means of several different collocation measures, pretty much all of them say "ok, yeah, those things are strongly attracted to each other." But of course we're most interested in ΔP, and the ΔP values here go as high as one, indicating that some of these things are really perfectly predictive of each other.

The more interesting thing though is to compare ΔP with bidirectional measures. And you will see if we do that, that more than a quarter of all these 2-word units are highly asymmetric, meaning that they actually only have an association in one direction, but not in both, which was pretty much all traditional measures would lead you to believe. And that tendency is not a function of frequency, but it's independent of that. So here if we take one ΔP value minus the other and just sort them, then we see that there are very many cases where there is a difference of at least 0.5 in one direction or the other. So 43+25 of all the occurrences I looked at and actually have a very highly asymmetric relationship to each other. And here, this is a graph showing that this is independent of frequency. So we have ΔP in one direction on this axis, we have ΔP in the other direction on this axis. Each of this little bubbles represents one bi-gram, and the size of the bubble represents the frequency. And you can see that there are a lot of words that have zero attraction in one direction but pretty high attraction value in another direction. But you can also see that there are big bubbles here, there are small bubbles here, I mean it's not like that's a function of frequency. Same thing here. There are a lot of bi-grams that have zero attraction this direction but a really strong attraction the other way round. And again, big bubbles, small bubbles all over the place, it's not just like what happens frequently in this way. And the little "X" here, that's *of course*. This great bubble where X stands for *of course*s. So really low association and really high in that.

Here are some examples, and I think these examples very clearly show why this is useful to actually keep this separate. So, in this panel, we have bi-grams where the second word strongly predicts the first one, but not the other way round. So here are some examples that I think make that very clear. So something like *old-fashioned*, the second word strongly predicts the first one but not the other way round. There is a ton of things that can happen after *old*, but if you have *fashioned* in a corpus I mean what other word would there be in front of it? I mean it's really hard to come up with something. *For instance* or *for example*, a lot of things happen after *for*. If you read *example*, that can be other things like *this is a good example*, but a lot of time will be *for example*. And then there are also nice cases of foreign language expressions where one word looks like an English word, but the other one doesn't. So something like *pot pourri*. So the first looks like English *pot*, and so that doesn't predict anything very strongly because a lot of things can happen after it, but *pouri* or whatever you pronounce that, there is not a lot of things in English corpus that might be in front of that. So it very clearly reflects these type of things.

Same thing the other direction. So here the first word is more predictive of the second. And here are some straightforward English examples, I mean what other things should happen after *according*, it's gotta be *to*, pretty much of the time, right? Or *instead of*, what else could happen after *instead* other than *of*? Not many different things would be there. And again, we have foreign language expressions, like *faux pas, gung ho*, my favorite is Italian *volte face*. And that is, *volte* predicts *face* really well, but *face* in terms of letters, that looks like *face*, OK? And a lot of things can happen before *face* in English. So this measure brings these things out really clear.

So this shows that ΔP can find this thing, I mean that ΔP can find stuff if there is stuff. Now, we also have to show however that it doesn't find stuff if there is nothing. So I took randomly-chosen word units. Basically, I took nearly 240 pseudo-randomly 2-word collocations from 8 different frequency bins to make sure there are no frequency effects that distort the picture and did the same type of analysis for those. We find that the average ΔP values are really close to zero as they should be if there is something going on. But as you can see, and I will talk about those in a moment, there are actually some words that have relatively ... some bi-grams of these have relatively high ΔP values. So we need to look at those to show that this doesn't undermine the usefulness of this value and I'll do that in the second.

This is the result in s plot though. It already shows that there are some that range really high, and some here. But we don't have this whole full-circle square type of thing that we had with the intentional bi-grams. But then the question is: What are these 8 bi-grams out of the 237 for which we got really high ΔP values?

So at first, I was annoyed to see that there were 8 that have these high values, but then if you look at them, it's OK because … I pseudo-randomly chose some, which of course means that some of the things that I chose randomly might actually be collocations. And it turns out that that is the case. If you look at the ones that are positive you get things like *I mean* and *I think*, which of course in a lot of discourse context, I mean those are just discourse markers in a way. And we have things like *I'm*, the way it is annotated in the BNC, I mean, that makes a lot of sense that that would be a bi-gram. And then *the faintest* and *the biggest*, which are possibly parts of collocations like *the faintest idea, I have the faintest idea* or *the biggest mistake* or something like that. In the other direction we have things like *sort of, lack of* and *cannot*. So the fact that ΔP does return some positive values as well because there are actually some collocations in this control group. So this is not a problem.

So we've seen that ΔP is more sensitive than the traditional type of association measures that is used because it has this sensitivity towards directional effects. It's not arbitrary. It's a well-motivated difference between percentages. It makes no distributional assumptions. It's just a difference of percentages, so you won't have to worry about normality and all that other stuff. And there were some experimental support. Both psychology and linguistic work that Nick Ellis and people have done. In a way, it's actually really interesting because a lot of corpus linguists try to validate some corpus findings against psycholinguistic data, using association measures often overlooking the fact that if you use a bidirectional measure on something that psycholinguistic is not bidirectional then of course you have to expect some sort of mismatch. So one thing one might to able to do now is to basically revisit a lot of previous corpus work with this type of measure and say "okay, it does provide for a better fit with the psycholinguistic gold standard."

So this all looks great, I think. But still what we are doing here is still a big simplification. Because so now we may compute this measure we basically say, there's *give* in the ditransitive versus *give* everywhere else. So all the other contexts is just one wastebasket category like *other*. We don't distinguish what happens there, so maybe this is actually not so great, and we should be more precise. So step number four is now the frequency of something in all its contexts, not just P versus whatever else, not just *give* in a ditransitive versus everything else, but *give* in a ditransitive, prepositional dative constructions, whatever in a particular corpus. So we wanna know how is the verb, the word, the construction whatever used in all the times? So this table now becomes this one. We don't just have *give* in the ditransitive, and *give* in the prepositional data, no. We list *give* wherever it shows up in phrasal verbs, in idioms and so on, which of course costs a lot more work. And in a way, this is just different types of dispersion. So it's not a dispersion of across corpus files or across corpus parts or registers or something like that—it's a dispersion across co-occurrence patterns. It's how is *give* distributed across the different grammatical contexts in this particular case.

Now, what is this relevant for? On the one hand, it's just descriptively more adequate. You don't just say, well, it shows up here and other stuff but you say what that other stuff really is, and that in turn can help, for instance, if you talk about these things in a language of the competition model it can help talk about the reliability of a form-function cue. We can see how strongly *give* is predictable of a ditransitive or some other of these constructions. It can identify cases of preemption so, is something very frequent in one context that it actually blocks out other usages of something? And also of course it has a lot of implication for learning and processing, because the more widely something is used, the more productive it is, the more likely children are to learn quickly that this is a productive construction and that it can be used in a variety of different contexts.

And there is some experimental evidence that shows just that. So for instance, Adele Goldberg and a former student of hers, they looked at how children and adults learn a new construction that doesn't exist in English, they just made up. And both the children and adults faced two different conditions. One is called skewed condition; one is called the balanced condition. Crucially, in both cases the learning experiment involved the adults and kids looking at 16 tokens, so 16 examples of that new construction with five different verb types so that was the same in both conditions. What was not the same is the frequency with which particular verb types were attested in the tokens. So in the skewed condition, there were five verb types, but one of them accounted for half of everything. In the balanced condition, the distribution was much more equal. And one way to quantify that distributional difference is by using a measure called relative entropy and relative entropy for this value here, for this distribution here is higher. And then what they found is that the skewed distribution, this one here, led to better learning than the other one. So it is important for us to know what these token frequencies are because Casenhiser and Goldberg (2005) show that actually makes a difference in terms of how quickly you pick these stuff up.

Boyd & Goldberg in a second series of this experiment show that pretty much the same type of logic applies to the learning of novel *a*-adjectives. And in general, we know that this type of distributional effects has a lot to do with how quickly we learn category information. So here is a relevant quote by Joan

Bybee, where she says: "in category learning a centered, or low variance, category is easier to learn". So that's something immediately bears on this and can be measured in this particular way. But we can only do this if we don't just say: ok, this so many times and then the rest some other times, but if we have the fine-grained frequency distribution for each of these types.

This also means that, on the whole, Zipfian distributions are probably very conducive to learning. What does that mean? So Zipfian distribution is a curve that usually looks like this. What it indicates is that a very small number of different types, let's say different verbs, different constructions, accounts for a huge share of all the tokens. And a large number of types occurs only very infrequently in a particular construction. If you look at that ditransitives, you will find *give* and *tell* and *show* and *send* in there a lot of times and those might already account for like 50% or 60% of all examples, but then you'll find dozens of verbs, which show up in a ditransitive only once or twice. However, it seems like a steep distribution of this type, and maybe the steeper the better, is something that very quickly helps children and adults, for instance, learning new constructions, new patterns, and new ways of using a verb for example. Specifically, if you look at Child Direct Speech (CDS), you find that verb types in constructions exhibit exactly that type of distribution, essentially helping children to recover what are productive slots in a particular construction. So for this type of stuff, corpus linguistics actually has a lot to offer because it is kind of us who are able to provide this type of distributional data, which can then be tested in a lab.

In the second language acquisition contexts, we have similar findings. Ellis and Ferreira-Junior look at sample different constructions in the second language acquisition of English. And they found that you pretty much get the exact same type of frequency distribution, and they find that, as I mentioned yesterday, the frequency with which learners are willing to take a verb and put it into a new construction, is strongly predicted by both frequency and association measures in the way we have discussed before. Thus, what we want basically is, like I said yesterday, we want to have much more comprehensive information, we want the complete frequency of something, we want the frequency distribution of how often do things show up there, we want to know how reliable is one thing in predicting something else. Ideally, we have all of this information for our corpus-linguistic analysis. Of course we are not done yet, because otherwise how simple would life be.

We're now adding another perspective namely the frequency of something in a variety of different contexts but now we return to registers and stuff like that in one corpus or corpus part as opposed to other corpus parts. How do we get to this basically by asking the question "what is the most frequent preposition of phrases in corpus linguistic papers?" It's something like: *in my corpus*. And you're always like "ok, but then, how does that generalize?" Because as you've seen a lot from the present perfect distributions. That is a huge simplification most of the time, because of the dispersion across files, because of the dispersion across registers and things like that. Now that's true of a simple frequency. If those already take a huge hit in terms of reliability and everything then you can imagine what happens once you run more complicated statistical stuff. So this table now unfortunately becomes this where we have the frequency of distribution of *give* across different constructions in one corpus and another corpus and maybe another corpus. Whoever said corpus linguistics is fun? So once we begin to look at stuff like this, a bottom-up strategy, the question of "to what degree can we combine those things" becomes pretty much indispensable because no one wants to read papers where you say "in this register" and "in that register" and "in this file", who cares. We want to generalize and we want to make abstractions and we want to come to more generic statements. So for that we will need analyses like this.

One question then becomes how do we do this? Here is one potential application which I think is really interesting. So I looked at the ditransitive and I looked at the ICE-GB, so the British Component of the International Corpus. I took all the ditransitives and looked at all the verbs that show up in the ditransitive and noted for each verb how much it likes to be in that construction. But I did it at the two levels of resolution you've seen before and the whole corpus. So once I didn't care the corpus has a substructure and then I took the five registers and then I took the twelve sub-registers. The result of this approach was a table, in this spreadsheet, that has 18 columns, namely, one column for the whole corpus 5 columns for the 5 registers, and 12 for the 12 sub-registers. And that was the columns and then it had 87 rows, one row for each verb that was attested. And each of the cells says, has a number in it said how much does this verb like to occur in this corpus part. How much does that verb like to occur in the same corpus part and so on? Obviously as you can imagine that table you can look at for as long as you live you will never see any patterns because I mean it's thousands of data points, very complex patterns, so you need some sort of statistical technique that helps you deal with this and one thing one can do on this is Principal Components Analysis. It doesn't matter how that works and what it does—it returns ways or suggestions how that table that has 18 columns can be condensed into a table that has fewer columns because the columns that are conflated are so similar to each other. So it basically says or this result basically says, you don't need to worry about 18 columns, just worry about four of them. If you look at those four you can actually recover more than pretty much 75% of the original table. So you take 18 columns, condense them into four but you preserve most of the information the original 18 one had. And of course four is easier to process than 18.

The question then becomes: what are those four columns? And this is interesting because it shows something that human analysts usually don't like to do. These four columns that the analysis returns, one reflected the spoken data, without private dialogue though, but all the other spoken parts. Then one was just spoken private dialogue nothing else, meaning that is different. Spoken private dialogue is different from everything else that is spoken, for ditransitives. And then written printed and then written unprinted. Now what does that mean? It means a corpus linguist who wants to look at whatever, ditransitive semantics, and who wants to be aware of what happens in different registers should distinguish these four corpus parts. They shouldn't take any of the levels that the corpus compilers intended but that's the levels that need to be distinguished if you want to be most precise about what the data show.

And the interesting thing here is that this is not just speaking versus writing. That's what pretty much every corpus linguist does all the time because obviously they are so different. The analysis here shows there are other levels that need to be distinguished. Secondly, it's not just a division into registers or into sub registers. The four principal components that come up here, they cut across different levels. And that's something human analysts don't want to do because you want to be tidy and you want to stay with one categorical system and stuff like that. Of course, the data wouldn't care. I mean if the data suggests a different level of organization and you want to stay true to those data, then maybe you will have to cut across these levels, too, and here we have a good reason because the statistical analysis suggests that that's what we should do. Did I mention that we're not done yet?

Number six. This is the last level of resolution to be added. And this is the scariest one. So this time we sort of zoom in a way by looking at the similarity of different uses of something in different contexts in a corpus. So far we've moved outwards from a particular frequency: we increased the scope by, sort of, more contexts more precisely, more contexts in more corpus parts, so now we sort of zoom in. So we looked at something like this. This is the schematic concordance of examples of *give*. Ok, you might find it in the corpus and then tabulated it. So this is the subject slot. This is the potential auxiliary slot, the recipient slot, if it is ditransitive construction. and the patient slot and so on. And basically what I want to say here it is that a lot of times we should also be concerned with what happens in between these different examples of, in this case, *give*. Why? First, we know that the similarity and analogy are relative notions anyway and I have shown you some examples of this before. And I think here I want to show especially with regard to the priming or persistence effect. I want to show you some examples that, like I said, are really scary for some reasons.

I want to distinguish two levels of similarity, one is *local* and you will see what that means in a moment. There is very interesting work by Benedikt Szmrecsanyi who looked at priming or what he calls persistence. He distinguished two different types of priming, two different ways in which what you heard before makes you use something again. And the first one is a kind of what everybody else mean when they talk about priming, that's sort of a default, the default is what he called α-persistence. A particular structure such as an active or a passive increases the probability of the same structure, active or passive, at a later point of time. That's the example I gave earlier when I said if I use a passive now you are more likely to describe something with a passive too, same structures. The scariest thing is β-persistence. A particular structure increases the probability of a similar structure at the next point where you have to make such a choice.

He did more case studies. I am going to report results of three here so he looked at the choice of analytic versus syntactic comparisons. So do you say *this is trickier* or do you say *this is more tricky*? Which of the two do you pick? Then he looked at future choices, *will* versus *going to*. Then he looked at particle replacement so the choice between *he picked up the book* and *he picked the book up*. Which of the two constructions do you use? And what he found is that comparison *more* lead to significantly more analytical comparisons. This would be something like *he likes this more than I do*. That would be comparison *more* the way he looked at it. And this structure *He likes this more than I do* makes people use analytic comparatives more, something like *This is more tricky than I thought*, although the structures are different. I mean *He likes this more than I do* doesn't have an adjective following but it still leads to an increase of analytic comparatives. Even worse, in a sense, *go* in the motion sense leads to more *going to* futures, which for many people doing grammaticalization would seem to be totally counter-intuitive but that's what he found. And finally, if you look at particle placement, if you have one phrasal verb and the next chance where you again have to decide which of the two word orders which you take—you're about to use the same phrasal verb—and priming is stronger. So lexical similarity, the presence of *more*, the presence of *go*, the presence of the same phrasal verb, even if it's not coupled with syntactic similarity, still results in the reactivation of structures.

The second example would be what I want to call global similarity and there is a very interesting paper by Neal Snider, he did that as a part of his thesis in Stanford where he looks at two things. First, kind of obviously by now, he looked at whether the repetition of verbs increases priming and it does just like previous work has shown. But then he looked at something that's more frightening and I will tell you why this is frightening in a moment. He looked at the overall similarity of one use of something to the potential next use or something facilitates priming. So he uses multiple features to quantify the similarity and a very nice similarity metric which I include here for you to admire. And he looked at the dative alternation, part of Joan Bresnan's dative alternation data set. And he did find indeed even if you control for everything else, the more similar the first use and potential second use are or would be more likely it is the people would use the same construction.

This has two implications: one is the priming of lexical material and structural material may be much more similar to each other because they both respond to similar factors. Second implication, like I said, this is now where it begins to be scary: Similarity operates on a lot of different levels or requires the analysis of inclusion a lot of different levels. But then this is the really bad part in a way. Things we count may be affected much more by previous things than by their properties themselves. What does that mean? It means that if you do analysis of the dative alternation, so you look at ditransitive and prepositional datives and then you do a concordance and then you look at this one example and the speaker did something that there is no way you can figure out why they did that, I mean, like the recipient is given, the recipient is a human and the whole thing denotes transfer and it's close spatial proximity. So everything in there says the speaker should have been using a ditransitive but they didn't. So you pour over the example for two hours and you don't know what they were doing. The answer might be: we'll go one minute back and see what that speaker did last time. A lot of completely unexplainable things, if you look at that example in isolation, might be just what happened a minute ago. What did the other speaker say in a completely different context, but still maybe that's a priming effect. So a lot of difficulties that we might sometimes face when it comes to predicting what speakers will do might just be things like that. That's why it's, first, generally important include the type of stuff. I mean that's why it is included in here. And what that also really shows is that doing a concordance of something and then picking a random subsample of this, that might be a very good way to waste a lot of time and shoot yourself in the foot. Because if you take a random subsample and do not consider things at least partially within their local context then you don't have that information that might actually be the main thing that explains the particularly intransigent choice at any point of time. So this is a sort of potentially really big problem.

What these all show is that people keep track of the distribution of information and they do that really fast. People can, within even short experiments, people can already pick up some patterns that result from the presentation of the experimental stimuli. And this happens really early. So infants are already doing this. Like I said more dangerously, this happens really fast. We find that in L1 acquisition with adult native speakers and in L2 acquisition in language contact situations and here are some examples I'm gonna go over really quickly because they don't bear on the argument so much. In L1 acquisition, 16 video clips were already enough to teach children a new particular type of construction. With adult native speakers I mentioned this study by Jeremy Boyd & dele Goldberg on adjectives, three exposures in the preemptive context and they got it. That is how fast this can happen. In L2 acquisition, we, a colleague and I did a sentence completion task, and over the course of the experiment, I think people saw 16 different items. Over the course of the experiment, there was a within-subject priming effect that made people more likely to use one construction over the other. In another study, we looked at the degree to which speakers of Turkish, Dutch-Turkish are affected, the degree to which people who speak Turkish-Turkish are affected by unconventional morphological patterns from Turkish speakers who had been living in Netherland for a long time. I think over the course of 8 stimuli alone they became more tolerant of what supposedly were unconventional utterances in Turkish. That was enough already. So this happens insanely fast.

Where does it lead us? I am gonna bring this up all together here. It leads to this a cline of co-occurrence complexity. Simple frequencies, simple frequencies and some contexts to an association measure, full cross-tabulation of something occurs or does not occur, plus adding dispersion both across files and across different construction patterns between what can be represented with a type of tabular growth design. We started with the table like this, where a word occurred in a construction 80 times, where another word occurred 60 times, a third word occurred 40 times. And then we increases it to something like this where we said the first word occurred 80 times, this is the same 80 as here. But now we add the information how often does it occur elsewhere. Then word two occurred 60 times and this is the same 60 as here. We just add how often does that word show up, which leads to something like this, namely, how often do all these words, and more, show up in this construction, and all the others. So again this is the 80 from here. And this 200 here that's everywhere else, the sum of this. We zoom into what that one frequency represents and show it in a larger context so that we can then compute, for example, the entropy of the distribution and see, does that have any implication for learning for speed processing, for speed recognition, all these kinds of things. I never said this would be a good-news talk. We can do one of these tables for each corpus or for each register because they will be different. Remember the example the other day? *Lift* will not be used intransitively in most corpus part but in the corpus part that is a body building online forum it would be. So that distribution will look different from corpus to corpus, from register to register. And again then, as I mentioned before pretty much, whenever we say word one or word two what we actually be more precise to also distinguish different senses.

Now to wrap this up so slowly, there is one crucial question here. And that is, if this is what should be doing, how can collostructions work at all? Because if this is what we should be doing, how is it possible that something like this yields any interesting results, which it does. So how is that possible? Well, on the one hand, I think it's possible because the association measure that is used is one that has some sort of cognitive reality, that's too strong. But it is a measure that has been designed to reflect things that happening in learning. And, of course, learning co-occurrence frequencies is what gives rise to this pattern. But there is another reason too. That is because it is a good approximation of the third approach, namely, this one here. The idea is that this mini table is actually a good approximation of this large table or a set of many tables.

Now how can that be? First because whatever percentage you observe, the measure weighs it by the frequency of occurrence. So high frequencies of occurrence will lead to a greater degree of statistical significance here. And secondly this is a little complicated maybe, but the token frequencies in cells B and C of these two by two tables approximate the token distribution of this huge table that you have seen. What does that mean? This is the big table. This is what we should be doing. Word one in construction one, in construction two, and in construction three, but then also word two in construction one and so on. So what I just say here cells b and c. What it means is cell b for this word is just the sum of all those. Cell a plus 200 elsewhere is 280. And c is just for word one, is just the sum of all those parts. And now the crucial thing is that for most of the time—This is d, the rest of the corpus. The little collostruction analysis of two-by-two table is an approximation of this more complex table just by summing everything up that is not that cell. That's why it works. Plus the other thing that helps is that all these token distributions will be Zipfian most of the time. So they will all have that shape and that makes the data comparable although they generalize or simplify things so much. This distribution will be Zipfian and then this is as well.

Now where does that lead us? Learnability of skewed input of the type that exhibits this type of curve may involve the type of cognitive anchoring as she says. But another way of looking at it is to say these type of curves just involve less uncertainty because the more types—uh, the more tokens fewer types account for, the lower the uncertainty, the better a child for instance is able to guess whether something can go there or not. One way to look at this, a Ph.D. student of mine talks about the whole thing include Hebbian of learning, for the interest of time I am gonna skip for now. I do want to make one final point though theoretical that is I think this type of perspective offers a very nice way to look at constructions or to define what a construction is.

In Adele Goldberg's first book, this is her definition of *construction*, she says it's a form-meaning pairing with at least one unpredictable property, something that doesn't follow from the other constructions we already know about or something like that. Then in 2006, she changed her mind and I think in a good way. She says something that is not predictable, doesn't, I mean, that's no longer a necessary condition. Even something is fully predictable, if it is frequent enough, it might still count as a construction. If you assume an exemplar-based view of the type I have been advertising here, where linguistic knowledge is sort of anchored or located in a multidimensional space, which has formal dimensions like syntactic co-occurrence information, which has functional dimension, which have to do with meaning, maybe information construction and other things, then there is no easier way to say this, a construction might be defined as something like this, namely, an uncertainty-reducing spike, of the distribution in some part of that space with at least one dimension being functional. Why does it have to be functional? That's because constructions are form-meaning pairings. There has to be some component that have to do with semantics, information structure, or discourse function or something like that. So that means a construction or children, for instance, recognize a construction when they realize that the distribution of some stuff is not arbitrary but helps reduce the uncertainty in predicting or interpreting what has just been said to them by their mother or by their caretaker or whatever, something like that. It's basically the kid realizing, Mummy doesn't use *give* randomly like all over the place; it always happens when she gives me something, when I receive something from her. The moment that is realized or hypothesized by the child the first time that reduces uncertainty because the next time a transfer situation might come up, or the other way round, the word *give* is used, the kid might expect "oh, didn't I just think she was going to do this?" And the two things co-occur together again then that reduces the uncertainty of the use of that word or the use of that construction or that co-occurrence. That will be a relatively testable operationalization of sufficient frequency. Because as much as I agree with Adele Goldberg, sufficient frequency is enough, it doesn't have to be unpredictable, I guess it's also obvious that *sufficient frequency* is not exactly a testable claim. I mean, we have to say what is sufficient frequency. Well, within this type of approach, we can say something is sufficiently frequent if it is frequent enough that it helps change the uncertainty distribution of one thing happening together with something else.

Last slide: if anything we need more complex tools and not simpler tools. We need to be able to test things against random baselines. We need to explore variability, dispersions of data, we need all sort of correlational structure to take into consideration. We need type frequencies, token frequencies, entropies, because I have shown you case studies that show that each of these things can be important and can be predictive in particular places. And all of that, in a sense, is relevant to collostructional analysis. But again, in all honesty, that is also a massive simplification. It works because it can make use of the fact that a lot of things are Zipfian and that simplification works, but in general, of course, we need more dimensionality, more different pieces of information. Because only if the quantitative tools that we are using are complex enough to handle complex data, we will actually be able to make the right type of analysis. Thank you.

LECTURE 8

# Bottom-up Methods in Cognitive and Corpus Linguistics: On Letting the Data Decide

This talk will be the first of basically the final section of three talks that make up the sort of coherent theme. In the last talk, I talked a little bit about the different types of statistical approaches and the kind of quantitative granularity if you will that I think one should assume one use in corpus data, and to some extent, in that talk, I early talked a little bit about the way in which sometimes the complexity and the multi-dimensionality of data makes it necessary, or at least recommendable, to use bottom-up approaches. Now in this talk today, I want to deal with this notion a little bit more, exemplify it in a variety of ways to show how statistical techniques that have been designed with an eye to reduce complexity, or to help explore complexity, can shed light on things otherwise difficult to notice.

In general, I mean, the field, I think, has become more statistical in two different ways. One is basically due to the fact that we have larger and larger samples of corpora right now, which of course is a good thing, and that also means whatever samples we get out of the corpus, I mean those become more and more complex both in terms of composition—so corpora have become more diverse, they incorporate more different register, more speakers and all sorts of things—and in terms of annotation—in the sense that we have, while, a few years ago, most corpora were not annotated at all or had some part of speech annotation, by now syntactic annotation, semantic annotation, and a lot of other features have become much more commonplace).

Secondly, we also have an increase in corpora that have some sort of temporal structure. So we have a lot of more historical corpora that show how something develops over time or over the last centuries, mostly, of course, on English but other languages are growing as well. We have a lot more language acquisition corpora where we can trace the development of things over time. And so that type of corpus sampling, basically as we'll see in the moment imposes some important restrictions on how we can analyze such corpus data, and I want to talk about that.

Now if we have corpora like this and if we look at a particular word or a particular construction, we annotate its uses for a set of features, and basically we end up with large multi-dimensional data sets not unlike the types I've mentioned before when we talk about behavioral profiles and stuff like that.

| doi 10.1163/9789004336223_009

For every item that is used in a particular context, we may have up to dozens of different features whose absence or presence we annotate for when we look at these examples. If that is the case then I think it's obvious by now that mere eye-balling of the data or thinking of what that might or might not show isn't really gonna do very much, simply because the human eye or the human brain isn't able to look at an Excel spreadsheet with 200,000 data points in it and see any meaningful patterns there. So we need statistical tools that help us handling this type of the stuff and there are basically two types of tools that are particularly useful or that basically any statistical methods can be classified into.

And in this talk I will talk about exploratory, or hypothesis-generating, tools, bottom-up methods that take as input a large data set where you have no idea what might be in there, or you have an idea but it is very difficult to systematize in a particular way, and the output of such methods would be some sort of revised way of looking at the data that imposes some structure, or suggests some structure. Now the other way to look at this would be with hypothesis-testing models, or hypothesis-testing statistics, most of the time that will be some sort of regression model, and that will be the topic of tomorrow morning's talk.

Now we have talked a little bit of hypothesis-generating approaches already, so if you remember from the behavioral profile talk we looked at, Dagmar Divjak and I looked in a study at nine different Russian verbs that mean 'to try' and we annotated these for a large number of features and then ran a cluster analysis on it, getting in return a dendrogram that shows three different clusters of verbs, where each of these clusters has a high within-cluster similarity, but then a large degree of dissimilarity to the other clusters. So basically, here, the input to the analysis was nine different verbs, and the output was, well, you can actually think about them as three clusters of three verbs each. That would be the most appropriate in the sense of bottom-up data derived classification of these.

The second example that we looked at very briefly was another example of behavioral profiles, we looked at a large number of senses of the English verb *to get*. We annotated them again for a ton of features, ran a cluster analysis on that, and we arrived at a tree diagram that, like this, grouped into several different clusters, several different senses to gather in clusters, because their distribution is similar and hopefully they also share some semantic or functional characteristics.

As I said before, however, a lot of times we want to apply bottom-up techniques to temporally- or sequentially-ordered data. And like I said before, these impose some restrictions especially when you look at historical data or if you look at language acquisition data. As a corpus linguistic of course the main idea is that you want to see how the frequency of something, a verb, a construction, a verb in the construction or whatever, changes over time, where, again, time might be centuries in a historical study or just two or three years as an infant acquires the first syntactic structure of his or her language. If you do that type of stuff, you'll have, I mean, there is a real host of problems that are really tricky and that don't necessarily shown up in many other corpus studies. For instance, the question is how we can find temporal structure in temporally-ordered data. One of these questions would be how do we make sure that if a frequency goes up a little bit or goes down a little bit sometimes, that there is actually a trend that covers a significant portion of the time.

Secondly, obviously historical data, I mean all things, constructions, verbs whatever over time, sometimes they will be more frequent, sometimes they will be less frequent—a lot of times that's gonna be nothing else, but, well, you have a different type of corpus sampling there. So how do we distinguish meaningful developments that were interesting for linguistic reasons from completely arbitrary fluctuations that we really do not care much about? How do we identify groups in temporally-ordered data? How do we, how are we able to, for instance in language acquisition data, distinguish different stages of acquisition? How does a child move from one particular level of proficiency or competence to the next, hopefully higher, level? And, finally, how do we deal with outliers or surprising data points? So how do we deal with data points that seem to go against the overall trend that we find in the data or that go against the significant subpart of the overall data? Just to make sure you can see this is not really, I mean, that I am not making these difficulties up, here is one particular example from language acquisition studies where we have on the *x*-axis, we have the age of a particular child ranging from about two years to four and a half years or something like that during the course of this data set. And on the *y*-axis we have the mean length of utterance of that child in morphemes. So obviously, as a child grows up, his or her utterances will become longer, will become complex, so kind of trivially on the whole, of course, the mean length of utterance goes up over time, I mean, anything else will be really weird. What is also clear, however, is that there is, apart from this overall trend, there are some real big fluctuations here, many of which might mean nothing linguistically, but that are just an artifact of whatever happened during that recording. So we would not want, like this big downward development, I mean, this probably doesn't mean anything cognitively, so we don't want to base a whole theory on language acquisition on the fact that here something different happened in the recording so suddenly everything went down. Even more extreme an example here, language change data, so again on the *x*-axis we have time from about 1420 to 1700, and on the *y*-axis we have the frequency of the current third person singular present tense form in English. So right now we use the -(*e*)*s*, so we say *give* → *gives*. Well, historically several hundred years ago the form would be *giveth*, so over the last 300/400 years that has changed, and again it's kinda obvious that over time, the frequency of the current form went up, I mean, duh, we know that, that's what we're using right now. Just as obviously there are some huge outliers if we look at these historical documents where some writer, for instance, obviously really didn't use the third person as much as many others in that time period did. And some people seem like they were ahead of the time, so how do we deal with these types of outliers, how do we make sure they don't mess up, whatever type of statistics we want to apply to these data later?

Now answering the first question on the basis of the corpus data is really easy, namely, "is it a trend, yes or no?" Here is a very small example on looking at the frequencies of *in*, just the preposition *in*, and the expression *just because* in the TIME corpus, which consists of approximately 100 million of words and ranges, and covers nine different time periods, essentially most of the 20th century. If you plot the frequencies of *in* in this case and *just because* in that case over these time periods, then it's kinda obvious to see that with *in* there doesn't seem to be in a lot of development, I mean people used *in* in the 1920s and they still do so. With *just because* it's very different, it seems like in the 1920s and 1940s there was not much use of that, and then that expression picked up until it arrived at a much higher frequency at the end of last century. So it seems like visually already this suggests that there is a trend for *just because* but not for *in*—how can we test this statistically? Essentially what that asking that question among asking whether there is a statistical correlation between the values of the time periods on the one hand and the frequencies of *in* or *just because* on the other hand. So basically you are asking, as the number of the time periods increases, does this number increase as well? And we've seen that for *in* it doesn't really do so at least not systematically, and with *just because* it seems to increase at least in different types of steps.

A lot of times people use a particular correlation coefficient for that, which is called Pearson's *r*. Unfortunately, a lot of time they actually shouldn't, because that measure suggests or has assumptions that the data usually don't meet, so an alternative that is better would be Kendall's $\tau$. If you compute that in these two cases, then for *in* you get a value of zero, a.k.a. no trend. If you do that for *just because*, you get a very really high correlation value that says, yes, as time goes by, the frequency of *just because* systematically increases. So very easy, I mean, to do this in R or any other software, it takes about 10 seconds, so that's easy to find out. So we'll see there is a trend between, let's say, the time period and the frequency of *just because*, but there is one problem here already, and that is that this type of method basically implies—even you may not want to imply that, but it does imply that—that the trend is sort of homogeneous and relatively monotonous or maybe even linear. But, of course, the data may not behave like that and in fact we have seen that the *just because* data, well, it first went horizontally pretty much and then only took off later, so it's not like there's one continuous super-regular trend. Sometimes the question might be—the question when you see there is one trend, the next question might be actually, is it one trend or are there different trends, are there different historical stages, and if so, how many, is it two, is it three, is it four? And even with a simple data set like this, this is already not obvious, I mean it doesn't get it—historically speaking, it doesn't get simpler than that—we have nine data points. But already if I ask people, ok, "how many different temporal stages do you see here?", it is not clear that we would all agree. Some people might say "it's just one trend." Some other people might say "well, there's this when nothing happens, and then there's this in the middle, and then there's this big stuff at the end. So that's three trends. Another person might say "there is this part when nothing happens, and there is a part when something happens. So there's two trends." So it's not clear how we would actually approach that issue in an objective way, and I want to suggest one possible way in which this can be done. And obviously once you ask how many different time periods or stages there are, the next question logically entailed by this is, how long are they? Is it 20 years, is it 50 years whatever? I mean, once you have that temporal sequence and you say it's three stages, then you still have different ways in which there could be three stages. And it is data sets like this that I think are very useful to basically support the call for exploratory or bottom-up statistical methods.

Here is one example about how this question of multiple trends can be addressed. It involves sequential data from Russian first language acquisition and it basically set out at the time to test the aspect hypothesis and the distributional hypothesis where the idea is that the way that Russian children acquire aspect—actually children in many languages acquire aspect—is that it's strongly coupled to the use with particular tense. So the aspect hypothesis specifically says something like you'll find a coupling of present tense and imperfective aspect on the one hand, and you'll find another coupling of past tense and perfective aspect on the other. And the way that cognitive-linguistic approaches or generally empiricist approaches would explain that is that the idea is that children form, I mean, on the basis of for instance verb islands or strongly recognized prototypes, children are at the beginning very inflexible and they just use everything, I mean if anything, they use what they hear exactly as they heard it, because in a way at the beginning they do not get that is exactly perfect ok to use imperfective aspect with past tense—that is possible, they just don't know at the beginning and then over time they release that very tight correlation as they notice, okay, actually mommy couples these things that I thought don't go together, I can say that, too. But the question is, when do they do that and how does that happen?

Now the data that we looked at for this particular case study were from my co-authors' corpus. We had 80 recordings from one Russian child from Sabine Stoll's corpus language of acquisition, which at that time included 6,800 utterances with a verb that were made by the child and approximately 32,000 utterances in the same recordings that contained verbs by all the caretakers. For each of these 80 recordings we coded the verbs for tense and aspect, and then drew up a table like this. So for every one of these recordings we had one such table, i.e. 80 such tables, so we basically count how often does a verb show up in past tense or non-past, and how often does it show up in perfective and imperfective aspect. And if we have a table like this then you can compute the expected frequencies, the ones expected by chance. And then you can compute a chi-squared test to see whether there is a correlation and to what degree is it rigid that children do prefer one tense with one aspect and the other tenses with the other aspect. If you run that on this table, you get a chi-squared value of this which happens to be significant, and you get an effect size of 0.33 which on the scale from zero to one indicates a relatively medium size, weak to intermediate type of effect. And you can do that for all 80 recordings and then see how do these values change over time? So what this value indicates specifically the higher it is, the more inflexible the child is, the less the kid uses one aspect with the other tense or something like that.

So this is the result for the children, for that particular child, and this is the result for the caretakers of that child in the same recordings. Now the question becomes, so what does that mean? It seems like there is not much going on over time for the adults, I mean, the values seem to hover around 0.4 without little change, so it seems like the adults know the language which is actually what we would want to find. And it also seems like for the kids, it's maybe going down a little bit like that. I mean there are some extremely high values here, but as the kid becomes older, you don't have similarly extreme high values here any more. There are some really low values in the right part of this panel but they are none here. So it seems like on the whole it's going down a little bit like this, which is what we would expect, namely a very inflexible coupling of tense and aspect at the beginning and then a reduction of that strong association as the child figures out "ok, I can actually do these other things."

Now what one might want to do then, if one were obsessed with inferential statistics, is you might do correlation or linear model and force a regression line through a data and compute a significance test. If you do that in this particular case, you'll get exact what you would want, at least at first, namely, you get a significant negative correlation. The fact that this value is negative indicates that as the child grows older, the Cramer's *V*-values go down, and it's highly significant so this doesn't happen by chance, so everything looks great. And if you do that for the adults, we can see that there is a negative slope, so the line does go down a little bit but not significantly so. So this is compatible with the assumption that the adults actually don't change that tense-aspect patterning any more, which again one would think makes sense. Now unfortunately, in a way at least, this is not the end of the story because statistics are very powerful, but only when you know what you're doing. And this is a nice example for, it is actually a nice warning for people wanting to use statistics before they know what they are doing. Just because there is another usage event of that—just because you can force a straight of regression line through the data, that doesn't mean that's in fact the most reasonable or most meaningful statistical analysis of this data.

At the time we looked at this and were suspicious, so we ran what is called a smoother through the data, so a regression line that, as you can see, is allowed to not be straight but that can have a curvature. That smoother shows a very different story, because at the age of three, it seems like the child does not have additional development any more, at that age it levels off. So it's not like it's one continuous trend all the time, it's a much quicker trend at the beginning and then at age three the child is at the level of the adults. If you do the same smoother for the adults, then you'll find there is a slight dip here in the middle but on the whole, I mean the adults end at the same level at where they begin no change. And again you can see that at the age of three the child is at the same height as the adults are the whole time.

So once you have that suspicion and that type of result, next thing you might do is what is called regression with breakpoints, so you basically decide "I need to fit two regression lines, not just one", and then in a way—how to do that is not relevant right now—but if one does it in that particular way, one finds that actually, yes, at age three, the first developmental phase ends, the correlation here is much stronger than it was before. And then at the age of three there is no correlation any more just like for the adults the whole time. So the actual developmental phases here are those two and not the straight regression line that you saw at the beginning. So what does that show? It shows that, apparently, the aspect hypothesis is on the right track, and apparently the distributional-bias hypothesis on the right track. That hypothesis states that the kids get this tense-aspect correlation from the input of the adults, and that—I didn't talk much about this—but this is also borne out by the data because note that the values for the adults, I mean they aren't zero, so the adults on a completely flexible, they don't couple whatever with whatever, they also have a preference for imperfective and present tense, and past tense and perfective, just not as inflexible as the kids are at the beginning. So both of these hypotheses are borne out and we find a differential pattern that suggests there are two different age groups or two different stages that need to be distinguished: one at the beginning until age three, rapid and highly significant adaptation towards what the adults are doing, and then one where there is not much development going on any more and no difference towards the adults. Crucial point though, simple plotting doesn't show you that, simple correlation, simple regression doesn't show you that, you need a more advanced exploratory type of approach that brings out these patterns in the data, otherwise you write the paper talking about one trend when they're in fact two or three or even more. And the bad news in a way then is that this is more frequent than you think. Here is a similar example involving historical data, namely the increase in frequency of *keep V-ing* construction in the TIME corpus. And here we're just looking at the last 15 / 20 years, and you can draw up a scatterplot that shows as time increases, so does the frequency of that construction per million words. And then each of these is the result for one particular year. And again if you bend over backwards to force a straight regression line through this, you'll get one and the correlation is high and significant and everything seems great. However, a regression with breakpoints suggests that the pattern is actually quite different, namely, the increase is not the whole time as this plot suggests, but really only until 1996 or 1997, and from then on we have some fluctuations but fluctuation that does not suggest an on-going trend or on-going development.

So now we've had two cases where basically we first saw, there seems to be an overall trend and then we used an additional method—in this case regression with breakpoints—to discover substructure within these temporal trends. But as I'm saying here, knowing there is one trend or several trends is often not enough, because it's not always obvious what the different trends would be or what the different substructures are. Now what people usually do when they want to find substructures in data is to use a cluster analysis type of approach, so an approach where we end up with a dendrogram that says this belongs together, that belongs together, and so on. Unfortunately, you can't really do that here. Why not? Because a cluster analysis would be—a normal type of cluster analysis—would be blind to the temporal ordering of the recordings. If we go back maybe to, back a little, here, [click the previous slide], let's take this one. So if you run a cluster analysis on this data set, then it's very likely that at some point at least of the analysis, the latest recording, when the kid is nearly five years old, gets grouped together with one of the earliest recordings, when the kid is two and a half years old, and that may make sense to a cluster analysis that doesn't know anything about cognitive development, but to any psycholinguist interested in language acquisition they would say "well, the kid is twice as old. I mean, he or she has made a lot of cognitive progresses and a lot of linguistic progresses. It makes no senses at all to pretend that the child at this stage has the same linguistic and cognitive system as a child at that age". A normal cluster analysis would do that, however, so we need a way in which we can basically prohibit, or do not allow, the cluster analysis to make these connections that span many years of language acquisition. The same would be true of historical data, if you remember—I'm not gonna scroll back to that—if you remember the increase of the third person singular as there were some values that really spiked up high or went-down low, you don't want to group that together with something that happened 200 years later, I mean English changed a lot during that time—you can't pretend that those are similar in some way. So here is an algorithm to the rescue, I called it variability-based neighbor clustering (VNC), which is a recursive algorithm and its main advantage for what we're gonna do here is that respects temporal ordering. It is a cluster analysis type of approach that only merges temporally adjacent stages so anything that a kid does at three years is only mergable with something that happens immediately afterwards or immediately before it but not with something that happened three years later.

This algorithm has a very nice pseudocode, which I have to brag about. It looks like this. If you are into programming, I mean, you should be able to implement it already. If you are not, then let me show it to you in a more intuitive way on the basis of, let's say, the TIME corpus data. So let's assume you have historical frequency data for 6 different stages: 1920, 1930, 1940, 1950, 1960, 1970, and then these are numbers, these might be occurrences per million words or something like that, any type of frequency information you get out of the corpus. So what this algorithm does is, it makes pairwise comparisons, it compares every time period to the next one. And once it has made all these pairwise comparisons it checks which of these two time periods are most similar to each other. And in this case it might find that 1960 and 1970 are most similar to each other. And then it merges those into a new time period and that new time period gets the name of the average of the two old ones. So now suddenly there is a time period *1965*, which contains the information from the previous two ones, And then, and now it's recursive then the whole thing starts over again. All time periods are compared to each other in a pairwise manner, but this time with the new time period 1965, the amalgam of 1960 and 1970. And again the algorithm looks for "ok, where is the highest degree of similarity", and it might find that is in the 1920s and 1930s, and then those get merged into something that is now 1925 and so on. So the algorithm basically successively takes time periods, merges those which are most similar to each other and proceeds.

If you apply that to *just because* data, then you get one of two types of results. This is the first type of results, it's ... For those of you who know what Principal Component Analysis or Factor Analysis is, this is a similar type of scree-plot—if you don't know what that is, the crucial point of this graph is to look where it levels off to the right. In this diagram it levels off at the number 3. So according to a particular criterion, whose technical details are not important right now, that means you should assume three clusters, three different time periods in the development of *just because*. And the second output of this algorithm then is the corresponding tree diagram. So this says there is an early stage in the development of *just because*, and that is the one when nothing much happened. Then there is a medium stage where *just because* picks up a little bit, and that covers the 1960s to the 1980s, and then there is a late stage that is 1990s and 2000s. According to the frequency information in the data that's the three temporal stages that one should assume. So now we have an objective answer to the question, how many, I mean, how does *just because* change over time? In this case, and the answer would be it does so in three stages.

So interim conclusions for that part, the correlational approach showed, just computing Kendall's τ or whatever it was in the beginning just shows there is a significant positive correlation. As time goes by, *just because* becomes more frequent. But as I've said maybe the data are not such a homogeneous data set, and one possible exploration, namely, this VNC approach suggests, there is a development: *just because* becomes more frequent over time but it comes in three stages and with two differently steep trends. By that I mean that this is a relatively small step, it just about doubles in frequency, maybe. But then this one is a huge step, that's like quadrupling in frequency or something like that. So we get very precise information about the number of temporal stages, their respective lengths, and the strength of the effects or the strength of the changes, the change of development during those time periods. And again, the nice thing here is that there is no bickering about who is right, I mean once you've decided to use the algorithm and have decided on operationalization of similarity and everything, just like with every other types of statistical analysis, then the results are at least objective, I mean, you can't really fake them in a particular direction that you like, I mean, you have to live with the results, which also means that sometimes this approach can help rectify decisions that researches have made which were incorrect. And the idea for this algorithm actually came when someone submitted a paper to my journal, and I didn't like what he did. So what he did basically is he looked at—this was the study I have mentioned twice before in the series of talks—it was the study that looked at the preferential patterns of *shall*+V over six different time periods, so this was Martin Hilpert's paper, where he tried to show what are the different patterns, what are different verbs that *shall* prefers to occur with over time. So he looked at a corpus that has 6 different time periods that were 70 years apart, so 1535, 1605, 1675, and so on. And then he did something that people do all the time in corpus linguistics, namely, they realize "oh, damn, my corpus is actually really small, so I'd better conflate some of the parts, because the I have larger frequencies and whatever stats I am doing will be more reliable." Perfectly valid decision, the problem is that he chose to conflate the data like this, he said "ok, I'm gonna make three different parts, namely, the first third, the second third, and the third third." So when that paper came, I was like "ok, maybe, but how do you know?" And of course at the time I didn't have the answer myself, so I thought what could one do, and I developed an algorithm and you have the result here already and that tree very clearly suggests that conflating these two is really not the best of ideas, because this actually goes with what happens later, and this goes with what happens earlier, so the real clustering should have been this one: the first half and the second half are not a division into thirds. Again, he made a decision that had a right motivation, namely, increase the sample size in each of the time period, but the way he did was not informed by statistical exploration of the data but by something that he considered practical. And basically, we ended up writing up a follow-up paper that discussed this type of result.

Another paper that we then did later shows another nice aspect of this type of approach. For this I want to return briefly to the development of the third person singular present tense in English. And this is the graph that you've seen before, so from 1400—whatever 1420—to 1700, or nearly 1700, obviously there is a rise of the current form, so this is the observed frequency of the third person singular -(*e*)*s* form. Over time, obviously, it goes up. And this is based on a pretty large data set actually, so we had about 21,000 cases altogether, 13,000 of the old form and then 7,500 of the new form. We use the Parsed Corpus of Early English Correspondence, which comes essentially in 233 time periods, where time period here is the noble scientific sounding word for 'this is the year in which a letter was written' that made it into the data base. As you can see, when you plot the proportions of the new form over time, there is an overall increasing trend but there is a ton of ugly outliers that make statistical analysis extremely difficult. The main point of this paper, which I will return to tomorrow, was to figure out, what are the things that sort of sped, or facilitated, that process over time? So, for instance, to what degree is there, I mean did this apply first to lexical or grammatical verbs? Were there articulating characteristics of the verbs that helps this along? And so on. So we basically approach this data set in their knowledge that at some point we would want to fit a regression through it to see which factors drives this change. Now if you ever try to fit a regression on something as ugly as this, you will know you'll fail because the regression will bend over backwards to try to make sense of these outliers, basically ruining all your other results in the process. So what we needed essentially was a way to systematically identify outliers, so that we have a motivation to discard them while at the same time protecting us from being accused of, "well, you deleted this one but not that one, because then your regression results are better". So we needed an objective algorithmic automatic procedure to be able to say "this one goes, this one goes, but this one will gotta leave in because it doesn't meet a particular statistical criterion."

So obviously we did a VNC analysis on this, and this is the result, this is the first result that we've got out of this. So again we have the time period on the *x*-axis, we have a particular difference coefficient on the *y*-axis, and an observed percentage of the new form on the second *y*-axis. And then you can see, obviously this is still a mess, I mean, it's very dirty, noisy, heterogeneous data, but you can also see that there are some cases, for instance, where a single letter or letters by a single person in a particular time period are completely different from everything around them. And the algorithm allows you, I mean, gives you a way to sort those out by degree of how much do they violate what happens in the data around them. So we adopted this stepwise procedure where the first time around, so this is the first result we got and these were ranked as highly extraordinary given everything else that happens around them in historical time. So we took them out, and then we ran the analysis again, so we get this tree, right? See, those are now gone, but we still have some outliers like, for instance, this one up here, which was marked as the next biggest violation of everything that is going on. So we took that one out, arriving at this particular structure, and then this structure is the one we actually worked with. So we assumed for the purposes of the analysis that follows that there were five different historical stages, 1, 2, 3, 4, 5. And the main reason for that actually was, this was the first paper that used that in this VNC approach in a regression type of context. You might ask, "why didn't you take out this one as well, because obviously this is very different from what happened before and what happened after?" The main reason for that is that we wanted to protect ourselves against the accusation that we cleaned the data so much that obviously that would be a historical effect. If you take this out, then you'll suddenly have a perfect increase from this, a little bit more, a little bit more, a little bit more, so people might have said, if you take out all the ugly stuff then of course you'll get nice regression results. So we left those in to say, even if you have stuff like that in there, you'll still get good results by assuming the stages that this algorithm returns. So I think it's a very powerful tool to apply to historical data, because you have a principled and you have a replicable way to determine the data points that are outliers and to maybe discard them. You have a bottom-up way to identify temporal stages, and funnily enough, in this particular data set the stages were actually better predictors of what happens if you take the years directly. So people who know some stuff about statistics would say that one should never use stages like this but just the years directly because obviously years are much more precise than the stages. However, we've tried that and actually the result were worse. So in this case the clustering yielded the better type of result.

Now another nice extension is that this type of approach, something that takes ordering or data-internal structures into consideration, can be applied to all sorts of things. Those might be temporal, and we've seen examples that are diachronic or a language acquisition data, but you can actually also run that on geographical data. And I want to give you a brief example of this, which hopefully will come out fine here ... maybe not. Let me just go with this. This is a map of UK. It comes with, I think 50 or something county-like locations, so here we have Oxford, Warwick, Nottingham, Leicester, or whatever—I don't know the UK that well—Banff, Kent, Cornwall, whatever, all these counties. And for all these counties, a colleague and I, we had syntactic or lexico-syntactic dialectology data. So he had a corpus where speakers from each of these regions produced data and then those data were annotated for the frequencies of particular grammatical characteristics or lexico-grammatical characteristics in their speech And now one might say "ok, wouldn't it be nice if one could show how different regions of UK differ from each other?" But again you cannot do that with normal clustering, because for some real statistical reason you might end up putting Somerset together with Inverness here at the top, which is like whatever 500 miles difference or something between or distance between those so that does really make a lot of sense, given the way we usually think about dialect continua, but you can tweak the VNC algorithm in a way that it respects geographical distance, so we don't just have temporal distance from one time period to the next—you now take geographical distance and only allow neighboring counties or districts to be merged. And if you do that, and now you have to watch very closely because the beginning results are hard to see, you see here at the top there is a little black thing appearing. So it's a recursive process, remember, right? Several time periods get merged in a recursive way, the same happens if you apply it to geographical data. So the two areas in the UK that are most similar with regard to the features we annotated is up there. So what the analysis does, it combines the two with a line, and sort of labels them with a number in the middle. The next one down here is . . . this is No. 2, so Wilmington, or whatever I don't know, and Somerset, I mean those are really similar to each other. The third one then is, I guess Cornwall and Devonshire or something like this, you can see that at the bottom, so that's the next most similar region. Then we have Kent and whatever that is as being merged similar, so the algorithm always just draws the line, I mean it finds out which of the two are more similar, it connects that geographical centers and then puts a line and the labeling in the middle. Then we have something here at the top, some connections here, and as we proceed you can see how more and more counties whatever are amalgamated, and you can also see that for instance until this point, the southern part and the middle part and the northern part, those are relatively clearly separated from each other, and only once you added whatever that is to the whole south, only then the whole south becomes similar enough to the middle of England to be merged. And then all these connections are solidified but note that this part here is still very very different, there is one relation here up the top similarity, but all of this here, which has now become really similar to each other, still does not connect all to the north, really different in terms of the syntactic features we annotated. So more and more and see now, this is now the first connection to what happens up here at the top, so you get a nice basically visual representation that shows how you can explore similarity relationships, in this case, between different geographical locations.

To wrap up, last slide. Obviously there is a lot of other tools available, so in a short talk like this, you can't discuss all the possible tools that there are. Multidimensional scaling would be one other exploratory approach that one might use where basically what you try to do is, you try to express the similarity of things on multi-dimensions by plotting things, by converting that into a two dimensional similarity plane. There is correspondence analysis, which is essentially a factor analysis on frequency data, and it has actually been used as an alternative to the Behavioral Profile approach that I talked about in the talk before. So Dylan Glynn (2010), for instance, from Leuven has applied that type of analysis (MCA) to the verb *bother* to figure out how these different senses are related, how different senses and how different syntactic patterns are related, so he finds agentive, agentive and a predictive *bother*, which have different syntactic preferences in British and American English. And a very nice study stretches the notion of what a corpus is, and Natalia Levshina applies this type of correspondence analysis to the semantic field of seating furniture, and the cool thing is that her corpus is basically pictorial, namely, pictures of different types of seating furniture in online furniture catalogs, which are then annotated for a variety characteristics such as “is there an armrest, yes or no?”, “is there an upholstery”, yes or no”? So, it is not a linguistic corpus when you have language in use, but it’s a corpus of pictures of something which gets treated and analyzed and interpreted then with the very same methods, so there is obviously a lot of potential here to extend this type of stuff to many different kinds of questions that ultimately then, for instance, relate to semantics or other linguistic subsequence. Thank you.

LECTURE 9

# The Use of Statistical Models in Cognitive Linguistics

Thank you very much for the nice introduction! What I want to do today is basically continue on a course I started yesterday afternoon, basically giving small a small subsequence of three talks that have to do with what types of different quantitative methods can be applied and how they—basically, what types of directions we have to go to find patterns in otherwise difficult-to-analyze data, and then the last talk today will basically conclude with how we can use the type of methods we talked about yesterday afternoon and this morning together with experimental data for validation.

So the first slide you have actually already seen, so I'll skip right to the bottom, again as yesterday the idea is that the data sets we look at especially from corpus linguistic data, they will become larger, they are becoming more complex, they are becoming more diverse and any type of analysis these days at least often leads to multi-dimensional data sets, i.e., data sets where we have, particular things we study, like the words, the construction, whatever, and we annotated or we analyze it in terms of various different characteristics that we usually add in a spread sheet and then hopefully later analyze in some particular way. If the number of features that you analyze in such study becomes large, then again the logic is that mere eye-balling of the data are just thinking really hard about the data, we are not gonna do it, so we will in such a study need some statistical tool, or tools, that help us to make sense of the data. Yesterday I talked about exploratory/ hypothesis-generating statistics—that might help with shedding some light on what happens to the data sets if you don't yet have a hypothesis but really just want to explore or generate a hypothesis as to what happens. Today then I will look at some applications that involve the statistical testing of hypotheses, i.e., you already have a hunch for what's going on with that data, but you want to figure out whether that hunch is correct or not.

Especially, in this connection, I want to talk very briefly about a very frequent dichotomy that people often bring up especially in order to position themselves as one or the other. And for some weird reason that is of course incomprehensible to myself, the importance of quantitative method is sometimes opposed or even questioned. There are those that argue that, for instance, many things in linguistics in general or cognitive linguistics in particular are

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_010

not really amenable to quantitative study, but that really qualitative analysis is all that needs to be done or even has some sort of deeper meaning to reveal that quantitative study cannot attain. But, and so the logic will be something like, since quantitative analysis, the analysis of something in terms of statistics, needs a qualitative analysis first, and then a qualitative interpretation anyway, why bother with the numbers? So how does qualitative analysis need quantitative analysis? Well, before you can do some stats, you have to have a spreadsheet where you have annotated sentences, or gestures or whatever in terms of some characteristics, so obviously you need to look at each example be that from a text corpus, a video corpus, from audio data or whatever, and mark in this spreadsheet, ok, this is a case where the subject is given, where the subject is animate, where the subject is three words long or something like that. In a sense that qualitative annotation or analysis pretty much always precedes quantitative analysis. Then, once you have the number out of stats you of course you want to interpret them, you do not wanna say "I have this great regression, guess what it means because I don't know." No, you have a discussion section that says, hopefully, what that means. So people would say "why bother with numbers in the first place?" Both of these views are wrong because on the one hand, qualitative analysis of any type needs quantitative analysis just as much as the other way around, if not more as I will talk about below. This is, first, because qualitative analysis usually requires or implies that you sit down and you label or annotate data points in a particular way. Right? You do a concordance of the verb *run*, you get 600 examples, and then you look at each of them and describe it in some way. You describe in such a way that you add some information in a spreadsheet that hopefully will later allow you to say something more general or more predictive about it. The point though is if you do that then basically what you end up with is frequency of occurrence, or frequencies of co-occurrence, of annotation. So if you have these 600 examples of *run* and you annotate the characteristics the subjects of *run* in terms of animate or inanimate or something like that, and you also annotate features of *run* in terms of are they transitive or are they intransitive, and so on, then you end up with co-occurrence frequencies such as: in this many cases, where the subject was animate, the sense was this, whatever the subject is given; in this many cases, where the object of *run* was inanimate, the subject was animate, and this means maybe something like that or this is the particular sense that refers to.

So we find things happen not at all, they happen maybe sometimes, they happen a lot of times, something may happen more than something else, or the other way round. But still at some point of time, following from a qualitative annotation process, you have some numbers. Again, if that data set becomes too large, just looking at it and hoping for inspiration, you'll have to do some statistical analysis in order to separate random variation from things that are linguistically actually relevant and meaningful. So in a sense then, even if you have a qualitative annotation step which I readily admit you always need, but you won't need the second step that makes that analysis intersubjective, so someone else may say "oh, yeah, there is that correlation;" it makes it replicable; and it makes it falsifiable and assigns it some predictive power. If you don't do that, if you just look at the data and think about something that seem to jump out from them, then you are likely to go wrong.

I want to show two brief examples, one basically using corpus data and the other one is slightly an unfair experiment involving some linguistics colleagues. The first one is concerned with this particular question, so let's assume in the week of September 11th in the US and the British press, you are interested in how the media coverage of the word *Muslim* has changed over time. You might be interested in that because, for instance, you have a hunch that over time since that particular event in time, the word *Muslim* is used with a growing number of negative overtones in press coverage, you might be worried about this, you might concern about what we can do educationally to counter that trend, something like that. Let's also assume that you do some sort of discourse-analytic approach, something like that: you do concordance of *Muslim* in press data, you annotate them for whether the word *Muslim* is used positively, neutrally, negatively, and then over time, let's say, you find there is highly significant super strong positive correlation that since 2000, as time goes by, the percentage of negative uses of the word *Muslim* goes up, like this. Those data are perfectly made up just to make that very clear. However, this is not gonna be enough because what you will need to really decide on what to say in your paper, you'll need what is a statistical interaction between WORD on the one hand, so WORD is an independent variable that covers words like *Muslim* and a few others that you will see in the moment. And secondly, TIME, so the years from 2000–2012. Now, why does one need that? Because if you talk about how *Muslim* has changed over time, then that is not particularly or not necessarily particularly meaningful unless you check what happens with other words in that semantic field over the same period of time. So the variable WORD right now only covers *Muslim* but it might cover other words in the semantic field such as the word *atheist*, and so this again totally made-up data, in this case the word *atheist*, the coverage of that word also has become a little bit more negative over time, but less so. Then, you might include the word *Buddhist*, and you check whether it has become used more negatively over time and the answer is "yes, it has." And then you might include *catholic* and "oh, my God, does that ever become negative?" And then you might include *evangelical*, and that has also become negative. So, ultimately, now we have a quantitative data set that started out from some discourse-analytic really qualitative annotation, but we now have quantitative data for several words and how their coverage and how their overtones change over time and looking at this is already too much, basically, to distinguish with a human eye alone whether any of these tendencies are significantly different or not. It certainly seems like *atheist* has changed least, and *Buddhist* maybe has also changed least, it seems like *catholic* has changed the most, look at these plots and some people might say “Nah, that’s kinda the same”, others might say “No, it’s not.” So you need statistics to answer the question whether these developments over time for the different words differ significantly or not. So even someone who starts out with totally qualitative question in order to make sense of what happens with different words in that semantic field, some point of time you have to do statistics or you just have to hope and guess your interpretation of these lines is correct or not.

Now one way one could do here is by fitting what looks complicated but is in fact quite simple is fitting what is called a linear model. So one tries to predict the proportion of negative elevations as a function of both the time axis and the different words that happen, and then what we find as the main results—obviously, I don’t want to discuss all of this here—but one thing that we find is that the variable TIME has a significant effect. So over time, all of the words in general have become associated, all these religious field words, have become associated with negative overtones. But then we also find, for instance, the word *evangelical* differs significantly from how *Muslim* has changed, and the word *catholic* does not differ significantly from how *Muslim* has changed. So statistically speaking, the answer is that red curve for *catholic* is not different from the black curve for *Muslim*, so both words have undergone the same type of negative overtone development. With *evangelical*, there is a difference in trend because the trend for *evangelical* is even steeper. And for *atheist*, this one here is completely significantly different from pretty much everything else. So, even again, even with something that is utterly qualitative in general and in terms of general research orientation, you have to do some sort of statistical analysis because you don’t want to write a paper that makes a big deal out of how *Muslim* has become used with more negative overtones all the time when in fact a lot of a religious terms have. I mean, that is a very different statement and it has a very different practical implications if you concerned with educational ways of addressing this or something like that. But if you just do the *Muslim* curve, and you don’t care about the stats anyway, then you’ll never see what the data actually have to offer.

Now, here is a second example, and again it is not quite a fair case study. It looks at the genitive alternation. So, the alternation between possessive "*'s* construction", *the presidents' speech*, and the often available alternative, the "*of* construction", *the speech of the president*. In a lot of cases, you can use either of the two constructions—the question basically is: can we determine with some degree of certainty which of the two patterns the speakers will use on any one occasion? So, I did a little experiment with some subjects who were tenured professors of linguistics and native speakers of English at the same time. And the experimental design was that they were given basically two tasks, or two questions. So, they were told I am interested in the genitive alternation and I'm interested in predicting which construction speakers would choose on which occasion or in which context. And then they were also told that I suspect that the following variables have an impact on what people will do, namely, the ANIMACY, the LENGTH, and the GIVENNESS of both the possessor and the possessed. Basically, saying, in *the president's speech*, the *president*, that is animate, so that would be coded as animate with this variable here, *the president*, that is two words long, and since *the president* is a definite NP, it's probably given in a particular discourse. If you say *the car of my mother* or something like that, *the car* is the possessed and it is inanimate, also has a length of two, and maybe was given, or something like that. We know from the literature these things have in effect but in those experiment I told the subjects that were what I was most interested in. And then they were asked to do two tasks: First, they were asked to provide generalizations as to how strongly these variables affect the choice of construction, so I want to know from them "what do you think how much does ANIMACY of the possessor and the possessed, how much does LENGTH, does GIVENNESS affect what's going on?" So, I was basically asking for what you statistically might call an effect size, how strong is any of those predictors? Secondly, I asked them for basically corpus frequencies, I asked them "what do you think which combinations of these things are very frequent with *of* and *'s*?" So basically, they could have said something that, and that would have made sense, they could have said something like, "ok, a lot of times with the *s*-genitive, the possessor will be animate and given, and the possessed will be inanimate, that would be the typical scenario that some human owns some concrete objects, and that will usually talk about using an *s*-genitive like *Peter's car, John's table, his book* or something like that. In addition, I collected some additional data, not from the subjects: I looked at a sample of corpus data which were coded for these variables. This is a pretty small sample, here; I just looked at 300 examples, namely, *of*- and *s*-genitives in equal frequencies, and equal frequencies of spoken and written data, so no big mystery here. And then I collected acceptability judgments from linguistically naïve native speakers of English, basically from a systematically-manipulated pseudo-randomized questionnaire I designed. The degree of GIVENNESS in the study was manipulated with a preceding context sentence: If a possessor was supposed to be given, and then the sentence the subjects were gonna judge had a sentence before that mentioned that possessor, so in that context, it was given.

So, what did the linguists say? They said yes, those things actually affect the construction, the choice of construction, which is good because that's true. Secondly, what's going on with regard to the effect sizes? For instance, they said possessors are more important for the choice of construction than the possessed, so in an expression like *John's table, John's book*, the fact that *John* has a particular characteristics, like being short, being animate, and maybe be given, that plays more role for what's going on with the *book*. And that's particularly true with regard to ANIMACY and LENGTH, not so much GIVENNESS according to what they said. The possesseds, on the other hand, they said, that's only relevant with regard to ANIMACY, so the fact that *the book* in *John's book* is inanimate that's important but its length is not. Apart from these two things, the answers were quite diverse; there wasn't much going on in terms of consistent patterns. Now in terms of frequent combinations, remember I asked them what thing happens frequently, they said, for *s-genitive*s, they focused on possessors. They would say "ok, *'s* genitives are really frequent when the possessor is short, given, and inanimate." Again, it makes a lot of sense as we will see in a moment. For some reason with *of*-genitive, they didn't talk much about the possessors at all, they talked about possesseds then, so about *book*, something like *the book of my father* or something, they were talking about—with the *of*-construction, they were talking about the possesseds, they would say those are long, new, animate, and abstract.

Now, what do other data show when we compared them to what the linguists said? So, in general, yes, these variables do have an impact and actually we know that from literature even before this little experiment. If you do a logistic regression on the corpus data, you get a really high classification accuracy. So, using these variables, ANIMACY, LENGTH, GIVENNESS, you can predict pretty well what subjects are gonna do. What about the effect size? Yes, indeed, if you look at the stats from the corpus data, you will find indeed possessors are more important than possesseds, that is correct. And yes, possesseds are important mostly when you talk about ANIMACY, and not so much about LENGTH and GIVENNESS. And ANIMACY in general is important, so with regard to all these things, linguists were quite on track. The length of the possessor, however, which they said was gonna be important is not important, neither in the corpus data nor in the experimental results, so, they were wrong there. But there's an even more important thing going on here than this little table of getting it right or getting it wrong suggests and this is the following:

All of the things the linguists come up with are just what statistically called main effects—they did not mention a single interaction between things. So, they talk about, ANIMACY is important, they talk about whatever, possessors are important, in terms of GIVENNESS and what not. None of the statements, none of the single one of them, was like "ANIMACY is important with possessors, but only if they are also given", or "LENGTH is important but only if things are inanimate." Not a single answer got to that second level of complexity, everybody was just talking about "this is what's happening everywhere;" none of the linguists said something like 'this happens but only in this subgroup of data, and in the other subgroup of data, it's the other way around', something like that. So basically leaving out all that interesting space.

Now, what about the frequencies of combinations of things? There were some guesstimates by the linguists regarding what is frequent or not. For the *s-genitive*s, they focused on possessors. For the *of*-genitive, they focused on the possesseds. I'm not sure this is really meaningful but it is interesting to note that it seems like they are talking about the first thing. In an *s-genitive*, you mention the possessors first and that's what the linguists are talking about. In the *of*-genitive, you focus on possesseds first, and that's the first NP and the second one, suddenly somehow does not get addressed any more although you would think, "well, why wouldn't that also be important?" So they would say things like this, "the possessors in *s-genitive*s tend to be short," which is not necessarily the case; "given," which sometimes is the case, sometimes not, and "animate," which is a correct prediction. So the question becomes, maybe the linguists are actually more led by their firm belief—those were cognitive linguists I should also add—maybe they were aided by the firm belief, things like short-before-long, and given-before-new, which are strong tendencies that the functional-cognitive linguists have talked a lot about. The same question applies when you look at what they talk about the *of*-genitives. So it seems like basically what they did is they focused on the first element, and then, I mean, I'm not sure this is what happens, but that's what it seems like, so they focus on the first element, because that is the one they talked about it. And when they talk about it, then whatever they say, takes LENGTH and GIVENNESS, and these two typologically very well-attested tendencies into consideration, and anything else, any interaction then were just dropped from the picture completely. Right again, not a single interaction was mentioned when the linguists were trying to guess what happens frequently.

Now if you look at the data and basically what that means, or if we look at the corpus data and compare them to what the linguists said, then here is one overview of some statistical results: So, on the *x*-axis, we have the length of the possessed, in *John's book*, that would be *book*, and *the president's speech*, that would be the *speech*, the thing that is the possessed, not the possessor, the *x*-axis starts with zero, but it begins with one and then goes up. On the *y*-axis, we have the probability that speakers would use *s-genitive*s, so they would say, *the president's speech*, not *the speech of the president*. So what the linguists were basically talking about is what is represented by these lines here, so this is the predicted probability of an *s-genitive* when the possessor is concrete, this is the predicted probability when the possessor is abstract, this is the predictive probability when the possessor is animate or human. This would be something like *the table, the flavor*, something like that, and *the president*. So what the linguists were basically doing is they were making one estimate for each of these categories, they would be saying, "Ok, when the possessor is concrete, then we don't use *s-genitive* very much." That's true, we don't say so much *the table's color*, or something like that, we say *the color of the table*, because *table* is inanimate.

Now, unfortunately, like I said before, they don't talk about interactions, but there are huge and significant interactions going on in the corpus data. So the linguists happen to be actually pretty good at estimating what happens with concrete things. So, the real results sort of cluster very closely around that blue lines, but they are totally off for abstract possessors and for animate possessors, which have conflicted tendencies that none of the linguists didn't even think about. Second result here, when we look at the animacy of the possessed, again, those are the types of main effects the linguists came up with when they thought about what might be going on, but in fact the empirical results are quite different and quite more complicated. So abstract and concrete things behave in such a way that their length effect reduces the *s*-genitive with the animates going up a little bit, if anything.

So, what does that show? Well, the overall estimates or overall effects were quite ok. So, I mean, the linguists noticed what would be important and they notice sort of in what sense, but again, they only provided estimates at a particular level of generality, namely, no interactions, thus missing out on all stuff that are really happening. If we look at the judgments concerning the effect size and combinations of variables, then some of them were correct, some of them were incorrect, all of them were monofactorial, and thus, grossly incomplete, because they missed out all interactional structure that both experiment and corpus data would show. So these types of things like the fact that LENGTH of possessed plays a role, but it is dependent on whether the possessor is animate or not, no one realized. Now, obviously, as I say, this is a slightly unfair experiment. Of course, any linguists would sit down probably for a longer time than I gave people the experiment. On the other hand, the fact that none of them even suggested the interaction is quite revealing as well. Second, quite open admission of course, what I've shown you here, I mean, is not a real cognitive linguistics analysis—the main point here to show is that if you rely on what you think might happen given what you know a little bit about the data or from some data set, then there's bound to be a lot of variability in the data that you are gonna miss. So ultimately, I think we will need to involve much more complicated statistical models, however terrifying that might sound to some. I want to show you one case I want to return to this example I talked about yesterday, namely, the development of the third person singular *s* in English as manifested in the Parsed Corpus of Early English Correspondence.

So again, this is the phenomenon we look at, and if you try to shed light on what happens there, what types of factors go hand in hand with that development, where does that change happen first, where does it happen later, there's a ton of things that might be correlated in different ways with this trend. And as usual if you look at the alternation or change phenomenon, they come from pretty much every level of linguistic analysis you can come up with. So phonological motivation might have to do with, in this case, might have to do with articulatory properties of the verbs, to which you then either attach this old form or the newer form. Syntactic motivations may have to do with, for instance, the difference between lexical and grammatical verbs. Semantic motivations, maybe verbs of a particular semantic class led the change compared to others that follow later. Sociolinguistic motivations, we know that a lot of times sociolinguistics changed initiated by women or speakers from particular socio-economic strata. Theoretically it's possible that the change arose in one dialect and spread from there to others, and there might be psycholinguistic motivations such as priming that has come up repeatedly now in some of my talks.

So, this is the data set, you probably remember the graph, so again we had about 21,000 cases, and we had this overall increasing trend with several of these outliers that basically needed to be omitted. In the last talk, I showed you how you can use this VNC analysis to have a principled way to decide which data points to delete. So remember we basically arrived at this 5-stage division of the historical data set where in general, we do have a sort of trend, of course, the new form takes over, but we also left in this one data point, this one, time period here with some letters that behaves a little bit odd, and I talked about it yesterday why we left that in.

Let's just go back here, this is a multi-factorial data set, I mean, we have a ton of features that might ultimately be correlated with how the change from the old form to the new form happened. We have 21,000 examples—please, no one believes that you can look at 21,000 examples in a spreadsheet annotated for 12 variables and see what's going on! So what we did is we fit what is called a generalized linear mixed effects model, a type of methods that currently is super hot in linguistics, maybe too hot bit I am not gonna talk about this in this talk. Statistically speaking, the dependent variable, so the thing that we try to predict, that we try to understand, is the variable called VARIANT, namely, the old one vs. the new one. And we have a bunch of independent variables or predictors. Those come into two classes, fixed effects, were the TIMEPERIOD as determined by the cluster analytic approach I talked about last time, the sex of the author: is the writer of that letter *male* or *female*? Second, is the recipient of the letter of the same sex, yes or no? Is there a priming effect so what did an author of a letter do last time when she/he had to make a choice for either one of the forms? Does the verb end in a sibilant, yes or no? Because, if it does, then one or the other form might be easier to pronounce, and yes, we are aware those are letters but then you'll see even in letters, even in writing, articulatory properties play a role. Second is, next one is the recipient close family member, yes or no? Then, is the sound after the *th* or the *s* form, is that a fricative, namely, either a *s* or a *th* or another. Well, again, the idea would be that maybe you avoid to have to use two *th* forms after another, try to avoid a knot in your tongue. Is a verb grammatical or lexical? And then all of these—that's a crucial point, very crucial point in fact because many of the historical study don't do that right—all these predictors are allowed to interact with TIMEPERIOD. That is really important, especially a lot of socio-linguistic studies still don't do that right. The reason for that is that, if you don't do that, and you get a significant effect of, let's say, final sibilant, let's take this one, grammatical verb, yes or no, so lexical verb. If you don't include the interaction of this variable with TIMEPERIOD then you might get a significant effect of grammatical vs. lexical, but you do not know whether it is constant over time. So you don't know whether in the 1400s, this variable had an effect that is different from what it did in the 1500s, 1600s, 1700s, which of course, if you want to do a diachronic study, may actually not be that great because the whole point is trying to detect change over time. So like I said, there are a lot of sociolinguistic studies that try to monitor something changes over time but they do not include that interaction, so they never know really whether they have a significant results or not. There's doctoral dissertations out there on historical change in the sociolinguistics that don't include that, which basically means they actually don't have data for the whole topic of their thesis.

Secondly, we included what are called random effects. These are adjustments to the overall regression equation that, in this particular case, help basically protect yourself against author-specific idiosyncrasies. Maybe a particular author never alternates—if that is the case; then you don't want that author's preference or total neglect of one particular variant to taint, or make more difficult, the analysis of the other data. So every author's name was entered into the analysis so that we can accommodates his or her idiosyncratic preference. And the same we did for verbs. If you look at the verb *make* for instance, then in the corpus as a whole, it is used with the old form 30% of the time—if you look at the verb *know*, it's used in the whole corpus with the old form more than twice as much. So verbs have strong preferences even over time, and this analysis is able to take this into consideration. So here is an example of the coding, just to give you a brief idea, we're obviously not gonna discuss this in great detail. This is the sentence from the corpus, the verb in question here is *promiseth*, and the dependent variable here, I mean, obviously, the speaker used the *th*-form, and then with regard to all the independent variables: This is an example from the 4th time period as determined by the cluster analysis; this has been written by a man to a woman; the last verbal choice was the *th* and that's actually in the same sentence—right we're looking at this verb, but there is another verb in the sentence in the third person singular and that form is also *th*—*promiseth* does end in a sibilant; the person, I mean, the woman that this man wrote to was not a close family member; and, the next phoneme is not a fricative. It is a vowel; and *promiseth*, is not a grammatical verb, it's a lexical verb. So, it's author-specific adjustment, so this was written by James Harrison; and the verb is *promise*. All of the 21,000 items were annotated in this way.

Then, we used what is called a stepwise model selection procedure. The idea here is to basically use a process that incorporates Occam's razor into the statistical analysis. So, the idea is that if something doesn't do anything in predicting which suffix the speaker will choose to write, then it is discarded because it doesn't add anything in terms of predictive power. The predictors that were discarded in this particular order, you can see all their *p* values are relatively high; they didn't cut it. And then summary statistics, those are not really that relevant at this point; the most important two things are down here. So in the model that has the random effects, in the model that was able to distinguish speaker specific and verb-specific preferences, of all the 21,000 items were actually able to predict correctly near 95%. At the time that we wrote this up, I had never seen that high classification accuracy so that was great. If you leave out the speaker and verb-specific adjustments, it's still a pretty good result: 86%. But you can see the difference that the two models have already made. Accounting for what speakers like to do, accounting for how verbs 'like to be used' boosts the classification accuracy considerably.

This is an overview of the main effects. So all the effects that do not consider TIMEPERIOD, or we will not consider how they change over time. This is the predicted probability of the old form. The high part of this plot, this is where people stick to the conservative and stick to the old form; the bottom part is where people are ready and happy to go with the new form. So, for instance, we can see that, yes, in general it's the female writers who go with the new form; it's the male writers who stick more conservatively to the old form. We also find an effect whether the recipient is of the same sex as the author, namely, when that's the case, so when a man writes to a man, or a woman writes to a woman, then in the data set as a whole, people used to stick to the old form whereas if you write to someone of the opposite sex, then you use the new one, which of course to some extent will be an effect of corpus sampling. Then these are the five time periods, we can see that, basically, the old periods 1, 3, and 2 stick with the old form and the new periods stick with the new forms of course. This is kind of what we find when we look at the main effects. But the crucial thing then is, like said before, actually, the main effects here really are not that interesting, because none of these things show how things change over time. Anyone of these numbers or statistics here says what happened in these 300 years altogether, but not what happened, I mean, how did things change over time? For that, we need to look at the interactions, I want to show you a few plots that basically highlight the types of results you can get from such analysis. In all of these plots, we have the time period on the *x*-axis. Period 1, Period 2, Period 3 and those are determined by the cluster analysis. All of these graphs on the *y*-axis, we have the predicted probability of the old form. If points are up here, that means people were predicted—authors, letter writers—were predicted to use the old form, if it is down here, like here, at the late stages, then people were predicted correctly 95% of time to use the new form. And then this is the interaction whether the recipient of the letter is of the same sex as the author. There are two possibilities whenever the author is of the same sex: there is a blue plus here in the data; whenever the author is of different sex, there is red minus here. The crucial point now is to basically track what these dashed lines do, which summarize the development of these factors over time, and the crucial thing here to notice is that if you ask such a question, "does it make a difference whether you write to the other gender yes or no?". Then it does, but not so much in the first two periods, I mean, the red and blue dashed lines, they are slightly different but not really very much. But in period 3 that is the first large, difference happens. So in period 3 you are more likely to be 'progressive' by switching to the newer form when you maybe want to impress the opposite sex, with your 'daring new' third person singular suffix. And then after that period, here is the biggest difference between the two, and when you write to someone else, you switch to the new form more often, but then after period 3, as you can see, the two lines are parallel and switch together in tandem to the new form. So the main effect that writing to someone of the opposite sex is in this time period, but after that, both basically addressee types switch together.

Here is another example, namely articulatory effect. Here, red and blue refer to whether there is a final sibilant in the verb, yes or no? So for *promise*, that would be "yes", for *give*, that would be "no". So here the question becomes, "does it make a difference whether a verb ends with a sibilant?" And again the answer is "yes"—for all of them "yes", because those are the significant interactions. But again you can see this time, periods 1, 2, and 3: the change occurs completely together. In those time periods this articulatory predictor made no difference. But then in period 4 both go down, but the red line goes down much further than the blue line, so the answer would be, especially in period 4, you are still likely to say *causeth* because that ends in a final sibilant, but you already say *comes* because it doesn't. So we can see which of the verbs, I mean, which types of verbs, already attract the new inflectional suffix and which don't.

One other example, another articulatory predictor is the phoneme that follows the third person singular suffix: is that a fricative, either *s* or *th* or is it something else? And again the articulatory thing begins to hit them in the 4th period, the three lines for *s*, for *th* and other, they moves along pretty parallel until period 3, where they are still all up there. And then, the red line stays up there so, when after the third person singular form, there is an *s*, people would still stick with the old form, but the other two develop together down to something really here at the bottom. So in period 4, you are quite likely to say *he sayeth so*, but you are unlikely to say *he sayeth that*, and I think me having to pronounce that already makes it clear why you would want to avoid that . . .

Final example, really quickly: "does it make a difference whether the verb is lexical or grammatical?" Again, time period 4 is one that marks the highest difference and the largest changes here. Especially in period 4, you begin to say *says*, because it's a lexical verb, but you do not yet say *doeth*, wrong inflectional ending there. It is only by looking at a data set of this size and this complexity *with* the interactions, that we can really tease apart what happens in what time period, to what degree does something make the change happen, in one class of verbs but not in the other class of verbs? I don't want to talk a lot about the random effects here because basically we don't have much to offer in terms of a story. This just goes to show that the pack of the authors—so, here I just over plotted all the names—as they show, it doesn't differ from the pack of the authors. But there are some people, here marked in blue that has very strong tendency to overuse one or the other form. And part of the reason why this regression model yielded the good result is that these people get statistically treated a little bit differently from the others in this type of model. The same is true of verbs, so most verbs here in black behave like most verbs, somewhat trivially, but some verbs have a tendency to go in one direction or with one form, some go in the other direction in the other form.

To sum up, if we allow for the idiosyncrasies of particular authors—remember the 2nd to last graph—and lexical items—remember the previous one, we can predict or classify really accurately which of the two inflectional suffixes will be chosen in any one point in time over the 300 years, which, I guess, suggests that we probably have the most important determinants of the alternation within our data set. Unfortunately, the people who make that corpus available don't make *all* the corpus data available, so they just keep the original or dialect data to themselves for reasons that are not quite clear. Same with socio-economic status, but nevertheless, if you get 95% out of 21,000 examples you gotta have something; that does not happen by accident. The overall conclusion is something like that you are more likely to use the new form when, obviously, you are born late (when you are born at the time when the new forms have already caught on); when you are a woman; when you try to impress the opposite sex; when you write to a man; when you use verbs without final sibilants like *come* but not *cause*; when you use a lexical verb, not a grammatical one; when you are primed or when you prime yourselves with having done that in the previous case -(*e*)*s*; and you use a word such as *the* or *that* after the verb so you don't have the clash of two *s*'s that would be difficult to pronounce. And, crucially, as we have seen, these effects are not constant across time: In some time periods, some of this things happen more—undergo a more radical change over different time periods.

Now, this was an example of relatively complicated statistical method if you want to do it right, and so people abstract from this that "regression analysis is really difficult, I don't want to do this, why don't I do my favorite test, which is a chi-squared test?" or something like that So, people always take away, this is massive overkill type of approach—do we really need such type of stuff, especially do we need to do, I mean, regression of this sort? In some sense, it's true there is a lot of simpler data sets, you don't have to throw a model like this and to figure out what is happening. However, sometimes it actually works the other way around. Sometimes you have a data set that looks treacherously simple, so people apply a treacherously simple technique to it incorrectly so, and then end up with conclusion they had better not put in writing. Here is an example from not cognitive but from corpus linguistics, a paper on second language acquisition. This is the table that was discussed in this paper, they (Laufer & Waldman) looked at V-N collocation, the V-N combinations, and whether they establish collocations, so things that have a particular high attraction to each other, along the lines I discussed earlier in this lecture series. They compared native English data from the LOCNESS corpus (Louvain Corpus of Native English Essays), and they then had three different levels of proficiency from the Israel Corpus of Writing English or something like that, namely, advanced, intermediate, and basic learners of English. To everybody who does not know statistics that well, they look at data like this, they go with chi-squared test. I mean I have 2-dimentional table with frequency data so why would I bother with anything that has the ugly *r*-word *regression* in it—when I can just run a chi-squared test. So what they do is they did eight different chi-square tests on different configurations of this table. For all of them, the results that they report are actually incorrectly computed, but that's a different story. So, they basically let themselves fool into assuming a particular degree of simplicity because the data look very simple, when in fact, that is not really warranted. If you do the overkill type of thing and run a regression on this data set, then you actually find that the intermediate—then this is what you find: On the *x*-axis, we have the corpora, so LOCNESS, the advanced learners, and beginner and intermediate learners together, and on the *y*-axis, we have the percentage of collocations, so the percentage of this out of that, this out of that, and so on. What the regression analysis shows, first, it shows the right results, not the wrong results. Second, it's one analysis and not eight. But third, it also shows that the real analysis would have to conclude an intermediate and beginners or basic learners are not significantly different. So you have to conflate them and treat them as one group, which is indicated here: LOCNESS, advanced learners, and then intermediate and beginners, they behave alike and significantly different from the rest. If you just do a chi-squared test, because you think "I just need to do a chi-squared test," you will never see that. Second, what you find is that this result is highly significant as on any of the chi-squared they do, but those are really only significant because they have a huge sample size. They have more than 40,000 items; pretty much anything you look at in corpus data for which it has a 40,000 data points is gonna be significant. What's more interesting is if you run a regression on the data, you will find the effect size is actually ridiculously low. In terms of what these results have, in terms of practical relevance these results have, I mean, that's nothing. It does not happen by chance, but what happens there is so small, is so tiny, so little predictive power that you probably wouldn't want to write it up. We had cases in the journal *Cognitive Linguistics* where a paper that has effect size like these, we just reject it, because you don't have anything. I mean it's significant but it does not say enough in terms of practical relevance, so, I mean, do it again, or submit it elsewhere, but we don't take it.

Last slide, basically one point to be made here is that using such methods as difficult and challenging as they seem to a beginner, but that is actually not a burden. I mean not only do you do it to protect your back against making generalization you shouldn't be making, but it's also an opportunity to see stuff in data you would otherwise not to find out. I talked about generalized linear models, and mixed-effects models that are now becoming more and more frequent. Again, I'm not completely on board with some of these developments yet, but on a whole, I think, it is a good thing to be able to tease apart what individual speakers or subjects do in an experiment, what happens with individual words or things like that, in general, it's a great thing. Other tools I'm interested to look at this point which have not been made into the mainstream are examples like this: Naïve discriminative learning, in a paper in a special issue of a journal Harald Baayen discussed, this is an alternative regression modeling that is cognitively more realistic because it is based on learning algorithms from cognitive psychology. As an answer to the question that was posed earlier by you, we have already briefly mentioned Bayesian networks and the good thing about them is that they force the research to make the hypothesis weigh more precisely than many regression models do that. So basically, we really, I think, we really should take a lot of steps and very decisive steps to make the discipline as a whole, make more use of these types of methods and others that the future might develop. Thank you.

LECTURE 10

# Corpus Data and Experimental Data: Examples and Applications

Thank you very much! In this last talk, I want to talk a little bit about the relationship between corpus and experimental data. The main logic will be something like, well, for nine talks now I have been telling you what to do, namely, things like "don't just use frequencies but also use dispersion measures or adjusted frequencies or something like that!" I told you "don't use just probabilities or co-occurrence frequencies or something like that, but use association measures such as the ones I listed here!" I told you "don't just use co-occurrence frequencies for instance or isolated examples taken from a corpus to describe the semantics of synonyms, antonyms, polysemous items and all that type of stuff, but rather use something that is much more involved like behavioral profiles!" I told you this morning "don't rely too much on introspective data to figure out why speakers are doing what they are doing or what they will do next, use multifactorial models!", and so on. In the course of these nine talks, I've sometimes alluded to experimental evidence or showed you like a little bit result slide or something like that, but in this talk I want to discuss these types of things in more detail, basically giving you, for these four different pieces of advice that I have talked about in the past week, give you some idea of how this could be experimentally tested and give you an idea of what type of results you get when you do that. Basically, we will revisit a few of the methods that I've talked about before, but this time more basically coming from the angle or from the question "what do experimental data have to add or do they validate or confirm the type of advice that I've given here that involve corpus data?"

For the first of these, let's remember basically that dispersion measures—statistics that quantify to what degree something is evenly or unevenly distributed in a corpus—help basically make frequency data more precise. So, I looked at data that show that words may be very similar in terms of the frequency with which they occur somewhere either in a corpus as a whole or in a particular register or something like that, but that they might still differ a lot in terms of that overall dispersion. This plot, you've seen before. We have the frequency of words on the *x*-axis, logged to the base of 10, we have a dispersion measure on the *y*-axis, and we can see that yes, there is an overall trend that more frequent words are more evenly dispersed, for especially medium-frequency range areas we have cases where words have pretty much the exact

© KONINKLIJKE BRILL NV, LEIDEN, 2017 | DOI 10.1163/9789004336223_011

same frequency but differ very widely in terms of how equally or unequally they show up in corpora.

This plot here, you have not seen so far: it compares for several thousand words—so each of these little dots here is one word—it compares for several thousand words their frequency and dispersion pretty much along the same logic here. And again you can see that, yes, there're overall trends that are captured here by these lines from different dispersion measures, but at the same time especially in the low- and middle-frequency areas, I mean, in the frequency of two, which means $10^2$, so 100, around 100, we can have words that are super evenly distributed with the dispersion value of 0.1, but we can also have words with the same frequency that have a dispersion value of 1, meaning they show up probably only in a single file of a corpus, but in that file they're super frequent. This was the type of result against which you want to basically insure yourself. In that talk I proposed a measure called *deviation of proportions DP* that helps safeguarding yourself against spurious frequency effects by providing this additional axis, and the logic was that at the time this measure has a lot of different attractive properties in the sense that it can handle differently large corpus parts, it's sensitive but not too sensitive, and all these other things.

One thing that as a corpus linguist, especially as a quantitative corpus linguist, you always have to protect yourself against is this idea this is just number crunching: so you find a different way of putting some numbers in there and getting some numbers out of there but it doesn't really mean anything. But as I've already mentioned before briefly at least there're a few studies that have shown that the addition of dispersion measure to the type of corpus statistic tool box that we have has substantial advantages. In the domain of second language acquisition, work by Nick Ellis has shown that dispersions have a higher degree of predictive power than frequency on its own. And in some work on my own, I looked at the degree to which dispersion measures correlate with psycholinguistic reference data. There's a master gold-standard data set by David Balota and Spieler from 1998 that contains reaction time, lexical decision reaction times, for a ton of words. You can basically download that list and see whether corpus frequencies or dispersion measures or both provide for a better fit, and if you do that then you'll find, and I'll show you a graph in a moment, that shows the dispersion measures actually in general do a slightly better job than frequencies and that also dispersion measures can differ a lot in terms of how good they are.

The same one can do with Harald Baayen's lexical decision task times, those are available as part of his statistics book. So you can just take those data, correlate them with different types of corpus statistics, and see what happens. And this is what happens when you do that type of stuff.

So here we have a plot that on the *x*-axis just lists of a large number of *d* for dispersion measures or *f* for adjusted frequencies, frequencies that downgrade or penalize words that have a particular frequency but that are very unevenly distributed. And then on the *y*-axis we have a correlational measure, namely the correlation of a particular measure with lexical decision times measured in milliseconds. The idea is that what you don't wanna see is that measure is pretty much in the middle because that says there is no correlation. So the corpus-based measures don't really do anything when it comes to predicting of psycholinguistic frequency [sic!] data.

Now if you take this data set and you now wonder, "where is frequency?", then it's here in the middle, approximately the third or fourth closest to zero, meaning nothing whereas dispersion measures such as the variation coefficient, the one that I proposed, or some down here are strongly positively or negatively correlated with the reaction times. So this graph alone already should provide good evidence, or should support the idea, that you don't just want to look at this one measure because, depending on what you're studying, you really end up with a very lousy correlation and bad predictive power. So there is really good reason to augment frequencies with dispersion measure, and the good thing about this is that actually also make psycholinguistic sense. This is a quote from a paper on language acquisition and you've seen this before "given a certain number of exposures to some stimulus, learning is always better when exposures or training trials are distributed over several sessions than when they are massed into one session." Exposure that's spread out as opposed to crammed into one particular session for example will always be better, and the way that we can quantify that, well, that is a dispersion measure. Again this is one source of experimental data that in this case strongly supports this idea of maybe adding a little bit of quantitative sophistication to the corpus data that we study.

Example No. 2 goes back to collostruction analysis. And again earlier I discussed in one whole talk some reasons or advantages of a collostructional approach to one that just looks at frequencies. And I did mention already a few studies that showed that this approach is often better than just frequency of co-occurrence. These were the studies that I mentioned on these earlier slides: one done by myself and two colleagues on using sentence completion data for the as-predicative, the second one by Daniel Wiechmann, who looked at correlated association measures with eye-tracking data, and the third one where we did the self-paced reading time study also on the *as*-predicative. But the thing is that, ideally, we would have more and more diverse evidence for this, so if the logic underlying this type of association measure approach is indeed correct, then we would hope that we can also find it in the linguistic systems of advanced learners, who may not speak a language as well as a native speaker, but if they are really advanced learners, their probabilistic language systems should be similar or should exhibit similar effects to what we would expect from a native speaker.

Here is a study where a colleague and I looked at *to* versus *ing* complementation, so basically the alternation between these two types of structures: *people began to make strenuous efforts* and *people began making strenuous efforts*. You can say both, but this alternation is quite tricky for learners, which of course has to do with the fact that these two constructions are semantically relatively similar but then also sometimes have surprising, I mean, for a learner maybe, surprising semantic differences. So here are some cases: *I remembered to fill out the form*: that might very well suggest you haven't done it yet, you remembered you would have to do it but you actually haven't done it, whereas if you take the *ing* form then that implies that you actually did it already. So sometimes there are clear semantic differences between the two but a lot of times, like in cases with *begin*, it's not necessarily clear what the semantic difference would be. What one does find, if you look at corpus data is that verbs so *began* or *try* or *remember*, they have very strong preferences to occur with either one of these two constructions, but then at the same time there's very little work on this from the second language acquisition perspective.

So what we did is basically a two- or three-step sequence of methods. First we did a corpus analysis of *to* versus *ing* based on British English native speaker data. The point of that was to basically find out what are these strong lexical associations. Are we able to do a collostructional analysis and see which verbs like to occur with a *to* construction and which like to occur with an *ing* construction?

And then secondly we did a two-type acceptability-judgment and sentence-completion task experiment and the design was actually relatively complex. I will try to show you that in a moment. The corpus analysis was pretty much straightforward. It proceeded along the lines that you've seen in Talk 5, we got a ranking of verbs to show up with particular constructions. Right now the actual ranking per se is not that important but we used the associations of these verbs with these constructions to design the experiment.

The experiment basically looked like this: it combined, so the acceptability-judgment task and the priming task were combined in one experiment, with the subjects not even necessarily knowing which of those two we were interested in, and of course it was both. So, subjects would read a prime sentence, something like *Sally tried to open the door*, which already contains one of these two constructions, in this case the *to*-construction. And then the subjects were asked to provide an acceptability-judgment grading for this sentence.

The subjects were advanced learners of English, native speakers of German. And then the second thing they were asked to do, I mean, the next thing they were asked to do is they were asked to complete the sentence fragment. So they got "John started _______", and then the only task they had was to complete that sentence in a grammatically acceptable English way.

And what we manipulated then were several things, namely, we manipulated the construction of this sentence, so in this case that would be a *to*-construction. But it could have been *Sally tried opening the door*, for example, in that case it would've been *ing*. Secondly, we manipulated, I mean, chose verbs in the slot that have different preferences. So in this case, *try* has a strong preference for *to*, as so here the sentence was a *to*-form and the verb also likes the *to*-forms more. If that prime sentence was *Sally tried opening the door*, then the construction was *ing*, but the verb like the other one better. And then the third thing that we manipulated was the preference of this verb, which here, I mean, that's the verb that likes the *ing* construction more. So there are basically three different potential influences that might result in speakers' doing something here. It could be something really local, namely what is that *this* verb likes? It could be something construction—I mean, they're supposed to fill in a sentence or verb phrase or something—it could be the whole verb phrase, so *try to open*, a *to*-construction, or it could be something very local but further away, namely, what *this* verb likes. And the question was what they would be doing.

We have 12 experimental items, six completions, six ratings, filler items, and pseudo randomization and all that other stuff and then the ratings here range from −3 to +3, depending on whether they thought the sentences were acceptable or not.

We got approximately 560 ratings from 94 subjects with these variables. We did a linear model to try to predict the acceptability-judgment rating on the basis of the construction of the prime and the collostruction preference of the prime. So the acceptability-judgment thing only looked at what happens here, namely, what construction is that, and what verb is that, and do they have the same preference? The expectation would be that if the verb here likes this construction, then the learners should like the sentence. If this verb likes the *ing* construction but we forced it into a *to*-pattern, then the learners should not like that *if* they have that sort of probabilistic knowledge that this analysis tries to quantify.

Now the good news is that the model is significant, highly significantly so, the bad news is the effect is quite weak. The second good news, however, is that it's exactly in the predicted direction. These two panels represented the same data just from different perspectives. What they show is that if the sentence they rated was a *to*-construction, then they gave positive judgments to verbs that also like the *to*-construction. That's positive. Whereas if they got a *to*- construction to rate, but there was a verb in that that was distinctive for *ing*, they didn't like that, so a very clear reflection—and the other way round, I mean, those are crossing—a very clear confirmation, I mean, weak effect but still a very clear confirmation of what the interaction and what collostruction analysis would predict for the behavior of the subjects.

Now what about the sentence completions? We got approximately, *exactly* 560 completions from the same number of subjects. And these included a number of *to*- and *ing*-constructions, and of course a lot of cases with neither of the two, I mean, the subjects got "He started ____", they could complete in any way, not necessarily using a *to*- or *ing*-construction at all. We analyzed that with a logistic regression. The dependent variable was "which completion did they choose: the *to*- or the *ing*-construction?" We only focused on those two possible options. And then we had the three independent variables that I mentioned earlier, namely, what was the construction of the prime? What is the verb in the prime, what does it like? And what's the collostructional preference of the fragment? So what would that be? And the question becomes, what do they do? The overall result returns out a very significant model with a pretty good classification accuracy. This value is supposed to be greater than 0.8, and it is. And, if you look at the individual effects you find that the collostructional preference of the verb and the target fragment, so the last thing they see before they complete the sentence, so the most local pertinent effect or factor, that is the strongest predictor of what the subjects would actually do. And it is in the direction that collostructional analysis would predict. So this is the predictive probability of a speaker completing the sentence fragment with a *to*-construction. And you can see that it's much much higher when the verb liked the *to*-construction. So this is a huge difference, highly significant, and exactly as predicted by the type of collostructional analysis. Again in this case we have two different types of experimental evidence, both acceptability judgments (and basically a meta-linguistic task) and the sentence completion that both show that the type of results that you get out of the corpus data when you do collostructional analysis are pretty good predictors of what subjects ultimately end up doing.

Third example, behavioral profiles. Earlier this week, I discussed the advantages of doing a behavioral profile analysis of highly polysemous items or multiple near-synonymous words. And in the discussion of these examples, I showed you that these, I hope, that these cluster analyses are quite revealing, they're quite versatile, it's a very flexible approach to be used. And this was one of the results that we saw when we looked at the nine Russian verbs meaning 'to try'. Now one problem: if you leave it at that, is there any additional evidence that supports this type of cluster-analytic result? The evidence we discussed when we did this first was relating semantic features of these verbs to previous lexicographic treatments in Russian and also for internal consistency with the result, but one problem essentially is that if you throw any data into a cluster analysis you will always get something, I mean, at the end of the day after a cluster analysis you will get a tree diagram.

So here is another tree diagram you might get for these nine Russian verbs. And that looks pretty good too, right? I mean, we have three very clearly delimited clusters: this time around one with four verbs, then these two, and then these three. The problem is in a sense that this is actually cluster analysis dendrogram based on random data, just random numbers, I mean, they mean nothing. There's no annotation, no function, nothing in there, and still you get a tree diagram that looks like "oh, yeah, cool! I have something to talk about", but in fact you don't. So, you need some sort of experimental validation to show that actually the nice-looking tree that comes out of there isn't just nice-looking in terms of "oh, yeah, I have structure", but also reflects something that goes beyond what you find if you just cluster anything like birthdays of your friends. So we did an experimental validation.

We had students from Moscow Computer Science and Economy Department sort sentences into groups. So they got cards, each nine cards with one sentence each on them, and the only difference was basically the verb meaning 'to try'. We actually did three experimental trials. I'm only going to talk about the last one; the results of the other two were identical in terms of implications. They were asked to put the nine sentences that only differ with regards to the verb in three groups of three verbs each, depending, or based, on overall semantic similarity. Every subject gives you back three stacks of cards that have three verbs each of them, and the question now is how do you analyze that? How do you compare that to a dendrogram or something like that? Unfortunately it's not that easy, but it's possible. We developed, or we used, two different approaches, and one of them is sort of newly developed evaluation metric that we came up with. The second one falls back on established statistics of cluster-analytic dendrogram comparison.

What we did is, in theory, and this is again it's not exactly easy and straightforward, but in theory the way it works is this. Step one is to generate a co-classification matrix which says, for each verb of the nine, how often it was put together into a cluster with each other verb. So for verb one, we get—all the subjects gave us the stacks back so we take the first stack, we take the first card and say "ok, this is *this* verb" and then we take the second card and say "ok, this first verb was put together with *this* verb one time", namely by this subject, and we noted that down. Then we take the third card and say "ok, this verb was put together with these two previous ones into the same cluster", so we noted that down. We do that for all the cards that all the subjects put into these stacks, which then means basically that you get a table that's—where high frequency say "ok, this verb many subjects put in the same group as this other verb". So you basically get a distance matrix. And I will show you that in a table in a moment.

And then we computed what are called the Pearson residuals that I mentioned before in an earlier talk. They're computed like this. The main point here again is that high values, high positive values, mean that something happens more often than expected by chance. And so ideally, of course, what we would want to see is that we get a lot of high values for verbs that ended up in the same clusters, because it means that subjects' sortings tend to replicate the clusters from the corpus data, which in turn will support the corpus analysis. We mark the highest Pearson residual in every row and then we scored, then we basically scored the subjects' performance or our performance by saying "if the verb that all the subjects like to put together with another verb, if that was in the same cluster as in the corpus analysis, then we scored one point, and if a subject put a verb into something most often where in a cluster that was different from the corpus analysis, then we didn't get a point."

So what does that look like in practice? We have a table like this with all the verbs in the columns, with all the verbs in the rows. And then this number here *a*, I mean, schematically represented as *a*, means that *a* subjects put *this* verb together with *that* verb in the same cluster when they sorted the cards. This number means, well, same here, so this number means, this verb was put together into the same group as this verb *b* times.

Then step two, we compute these Pearson residuals. These are the actual data and I've arranged the verbs in the order of the clusters. So this was the first cluster, this is the second one, this is the third one, which means we want to see basically that there are positive values in the yellow things where people put *this* together with *that*. So in this case, in every row, the highest residual is marked in blue. And so this is very nice because it shows that this verb was put together most often with this one and in fact they are on the same corpora cluster, same here, this one, no: so people put this verb most often together with that one, but that's not what the corpus analysis would have suggested. But then the other blue numbers are all in the nice yellow triangles like we want them to be. So we get eight points. Eight verbs behave the right way and *silit'sja* doesn't. As usual that actually doesn't, I mean, ok, eight out of nine kinda speaks for itself, but theoretically we do need a baseline because sometimes if you have monkeys sort nine sentences into three stacks each, some of them will do this, so we need a random baseline.

The fourth step was basically, I'll skip the technical details here, the fourth step was comparing the values that we had, eight—we basically figured out how good is that, given the fact that we could have minimally obtained zero and maximally nine, and again, I mean, eight out of nine already makes a strong case, but we still need a baseline comparison and we did that in two ways. First, just using probability theory, second, using a Monte Carlo simulation. It does not really matter what that is right now; it's sort of testing what would you get randomly and then see how much better than that is one.

The Monte Carlo simulation, we did that sampling 100, 000 times and yes you do that with programming, not manually. And then the *p*-value that we get is this: So that means getting eight out of nine right by chance is extremely unlikely, that it's possible but it usually doesn't happen, which means the fact that the subjects sorted the sentences in a way that is very very compatible with the corpus data is not random. It says something about the method probably being a good representation of the speakers' probabilistic knowledge. This is a different way of representing these data.

Second type of verification, we had sorting data from the subjects. Every subject says "this is this similar to this", "this is, (this) similar to that" . . . Actually, we could compare the results of the cluster analysis from the corpus data—let me bring this back here, right, this one—so, we could take the dendrogram from the corpus analysis, and then compare it to a dendrogram that we get from the sorting data. So we computed the cluster analysis on the sorting data. We get this tree diagram. The tree diagram suggests, this diagnostic suggests that there are in fact three clusters, too. You can already see that this is not the same solution: the corpus-based solution had three, three, three verbs—this one has four three two, so there are differences. But if you compute a particular statistic that says how good or how similar are two dendrograms to each other, then this reaches a pretty high value indicating that there's a good overlap, that the subjects' sorting is on the whole very compatible with what the corpus analysis suggested. Again, both kinds of analyzing the empirical, I mean, the experimental data result in this case a relatively clear and statistically significant confirmation of the corpus-based analysis.

Final example, I discussed earlier today in fact that multifactorial statistic modeling is very important when it comes to studying multidimensional complex corpus data. But one thing that's definitely already problematic is that the mathematics underlying many of these models is hardly ever cognitively realistic. I mean, a regression modeling approach works on the basis of observing least-squared summed deviations which is something we in our heads don't do. So it would be nice if we had a better way to determine whether what they predict doesn't just have a good classification accuracy in terms of how well does it do for the corpus data, but if it was also able to predict experimental behavior. In a way that's comparable to the *to*-versus-*ing* study earlier. That study looked at verb-construction associations—what I wanna look at now is more cognitively or cognitive-linguistically relevant notion, namely, that of the prototypicality of instances of constructions. In this case, the target of study is the dative alternation in English, so the very well-known constituent order alternation represented here, so the difference between *John gave Mary the book* and *John gave the book to Mary*. Again most of the time, you can use either of the two constructions and pretty much all of the time speakers have no clue why they do which. And again, as before, there is a large number of factors that determine this alternation having to do with phonology, syntax, word etymology, it seems, discourse-functional factors that have to do with givenness and newness, length of the patient and the recipient, all sorts of things. And semantic factors too, everything.

Now in an earlier study, I coded a variety of these features in an actually ridiculously small corpus sample. One of the things I coded was whether the verb phrase *gave Mary the book* denotes transfer, "yes" or "no". So in this case it would be coded as "yes". I mean, *gave the book to Mary*, that does mean that the book changes possession. I coded whether the patient and the recipient were animate, the noun phrase type of the patient and the recipient, whether both of these were definite or not, how long they were—I think I used a number of words at that time—and to code for something such as givenness, so how much is the referent of the patient and recipient inferable or even totally given in that particular context, I counted in a preceding ten clauses how often the patient and the recipient were mentioned, which is in part the explanation for why the data sample is so small, because that's a ton of work. Because for every corpus example, you have to read ten clauses back and see how often does something show up. And I also coded the distance to last mention of the patient and recipient if there was one. So the idea being that if something was mentioned one time in the last ten clauses, then that could either be the immediately preceding one in which case it would be highly given, or it could be one time but that was nine clauses away, so that's already a big difference. Times of preceding mention and distance to last mention are correlated but not perfectly so, so I made sure that I covered both of these different dimensions of potential givenness in the corpus. Then the two main questions at the time: first, can we predict the constructional choice that speakers will make? But then also, can we identify on the basis of what the corpus data show which of these instances of the constructions in the corpus that I coded for, which of those are close to what might be considered the prototypical ditransitive or the prototypical prepositional dative?

Now with regard to the first question, I used what is called the linear discriminant analysis at the time—a logistic regression returns pretty much the same result—and the model is highly significant, and it has a very high classification accuracy, so nearly 90% of the speaker choices were classified correctly. How does this work? It's actually not really that different from a logistic regression: The model or this type of analysis computes on the basis of the data in ways that are irrelevant right now what is called the discriminant score. And then the exact number of that discriminant score is used to predict with different degrees of certainty what a speaker will do.

In this particular case, what happens is if the discriminant score was greater than zero, then the model predicted that the speaker would use a ditransitive construction; if that score was less than zero, then the model predicted the opposite, which also means if that score was very very close to zero, then that kind of means the model says both sentences would be ok, you could use either. S speaker shouldn't have a strong preference for either of the two, because the model doesn't lean neither way particularly strongly. The further away the score is from zero, the more that particular corpus sentence has the characteristics that are typical for one of the two constructions, and, because of that, the more certain the prediction is that the statistical model arrives at.

If you do that, and you look at the highest and the lowest discriminant scores, so the sentences for which the model predicts "ditransitive" most strongly and the sentences for which the model says "this *has* to be a prepositional dative, no way, you can use a ditransitive there," then these are the sentences that you get. For the ditransitive, you get something that is pretty close to transfer, metaphorical transfer if you will: very nice the recipient is a personal pronoun, it's super short, and then the patient is longer, it's less given because it's something abstract, not something concrete who is a participant, so this sentence contains a variety of characteristics or exhibits a variety of characteristics that are really highly associated with ditransitives.

The example most prototypical for the prepositional dative according to this analysis is this one. The sequence of names doesn't matter right now, and then *gave a new impetus both to the study of these themes and to the action upon them*, so a superlong *to*-prepositional phrase which, given short-before-long in English is put at the end, again a really good match with what we know from previous literature, prototypical caused-motion constructions might be like, but again with something abstract in the patient phrase.

Now if those are the prototypes, or sentences that are close to prototypes, from that it follows that people should like these sentences if they are shown in the construction with which they are actually attested. If a sentence gets a high score for a prepositional dative, then speakers really shouldn't like it if you change that to a ditransitive, because the sentence has all the characteristics of a prepositional dative, so why would it be uttered in the wrong syntactic construction, and the same of course works in the other way round. To test whether the corpus analysis actually could do that, I mean, was precise enough for something like that, I did a little experiment.

The first independent variable was prediction. So I picked two sentences that had really high ditransitive scores, two sentences that had really high prepositional dative scores, and two sentences whose scores were really close to zero, meaning speakers should like them either way. Second independent variable, I took these six sentences out of the corpus, they came in one construction, but I also changed them to the other one, the one that speakers usually would not be expected to like. And then the dependent variable was the acceptability judgment again ranging from –3 to +3.

36 native speakers of English participated in that experiment and there were all the usual experimental controls: filler sentences, randomization, a ton of other distract sentences, and everything. In fact the distract sentences were two other experiments that were run at the same time. And then the predication was that the speakers should like the stimuli that they were given, when these stimuli were presented in the structure in which they were attested in the corpus and which should be the preferred ones.

As usual, a linear model, this is why regression should be your friend. The model is highly significant. The effect is not that strong, but still. But then the question really is, "what happens with the predicted interaction?" The strongest effect in the whole linear model is exactly as predicted in fact.

So here we have a case where the prediction that was made on the basis of the corpus data is on the *x*-axis. The corpus data might say, "something is really strongly ditransitive." The model might say, "something is a strong prepositional dative", or it might say "this sentence, doesn't matter what you do". And then look what happens. So, if the corpus prediction is "prepositional dative" and they get a prepositional dative, so the red result, then they like that, positive judgments. If the prediction is "prepositional dative", but they get the sentence in the ditransitive syntax, blue, then they don't like it, less than zero. And the opposite here and, in fact, I didn't even expect that it would be that great, but when there's no preference, they like it either, both of them positive. The predictions you would expect on the basis from the corpus data were exactly borne out as the slopes here of these curves suggest.

To sum up, for many of these tools or methods or whatever that I've discussed here over the course of the week, you've seen now sometimes several different ways of gathering experimental support or validation data for this and ideally of course we would always seek this type of converging evidence. I mean I love corpora, but it's not like that they can do everything. They're really nice in the sense that you can ask them to do stuff for you at 2 am in the morning when no subject ever would show up. But you do need additional data and confirmation from other sources as well so it should go both ways. Ideally we would do that with different methods with different data sets in each of these types of data. You've seen cases now where we did judgment, acceptability-judgment experiments, sentence sorting, sentence completion, all sorts of other things like that. So ideally the more diverse the data sets that you have from each of these two major paradigms, observational and experimental data, the better.

Now the downside obviously is that it's ton of work. And also it's not without its own problems, because mapping or making sure that you have comparable data from the observational and the experimental side is not always easy. I mean, frequencies in corpora will never perfectly predict lexical decision reaction time: there's too many other things that happen as well. So sometimes making that step is actually quite tricky. But only if you do that can we make sure that we actually make real progress with how we analyze the hypothesis that we have. And only that type of step guarantees that the way that we do cognitive linguistics is empirical and is actually interdisciplinary rather than us sticking to ideas about what type of things might happen in speaker's minds—if we can test those on the base of different sets of data, the result will ultimately always be a little bit more reliable than if you don't. Thank you!

# References

Aarts, J. 2002. Does corpus linguistics exist? Some old and new issues. In L.E. Breivik & A. Hasselgren (eds.), *From the COLT's Mouth . . . and Others': Language Corpora Studies in Honour of Anna-Brita Stenström*, Amsterdam: Rodopi, 1–19.

Albright, A. & B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90. 119–161.

Ambridge, B., A.L. Theakston, E.V.M. Lieven, & M. Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development* 21. 174–193.

Anderson, J.R. 1982. Acquisition of cognitive skill. *Psychological Review* 89. 369–406.

Arnon, I. & N. Snider. 2010. More than words: frequency effects for multi-word phrases. *Journal of Memory and Language* 62. 67–82.

Arppe, A. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography—a study of synonymy. Unpublished Ph.D. Dissertation, University of Helsinki.

Arrpe, A. & J. Järvikivi. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3. 131–159.

Atkins, B.T.S. 1987. Semantic ID tags: corpus evidence for dictionary senses. *Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary*, 17–36.

Atkins, B.T.S. & B. Levin. 1995. Building on a corpus: a linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8. 85–114.

Baayen, R.H., R, Piepenbrock, & L. Gulikers (ed.). 1995. *The CELEX Lexical Database* (*CD-ROM*). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R.H. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5. 436–461.

Baayen, R.H. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11. 295–328.

Bauer, L. 1983. *English Word-formation*. Cambridge: C.U.P.

Bell, A., J.M. Brenier, M. Gregory, C. Girand, & D. Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60. 92–111.

Berg, T. 1998. *Linguistic structure and change: an explanation from language processing*. Oxford: O.U.P.

Bergen, Benjamin J. 2004. On the psychological reality of phonaesthemes. Language 80 (2). 290–311.

Biber, D. (1993). Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics* 19. 549–556.

Biber, D. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14. 275–311.

Biber, D., S. Conrad, & R. Reppen. 1998. Corpus linguistics: investigating language structure and use. Cambridge: C.U.P.

Bolinger, D.L. 1968. Entailment and the meaning of structures. *Glossa* 2. 119–127.

Bowker, L. & J. Pearson. 2002. *Working with specialized language: a practical guide to using corpora*. London: Routledge.

Boyd, J.K. & A.E. Goldberg. 2011. Learning what not to say: the role of statistical preemption and categorization in a-adjective production. *Language* 81. 1–29.

Bresnan, J., A. Cueni, T. Nikitina, & R.H. Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Krämer, & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Brook O'Donnell, M. 2011. The adjusted frequency list: a method to produce cluster-sensitive frequency lists. *ICAME Journal* 35. 135–169.

Butler, C.S. 2004. Corpus studies and functional linguistic theories. *Functions of Language* 11. 147–186.

Bybee, J.L. & J. Scheibman. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37. 575–596.

Bybee, J.L. & S.A. Thompson. 1997. Three frequency effects in syntax. *Berkeley Linguistics Society* 23. 65–85.

Bybee, J.L. 1985. *Morphology: a study of the relation between meaning and form.* Amsterdam & Philadelphia: John Benjamins.

Bybee, J.L. & D. Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. *Language* 82. 323–355.

Bybee, J.L. 2010. *Language, usage, and cognition*. Cambridge: C.U.Press.

Cannon, G. 1986. *Blends in English word-formation. Linguistics* 24. 725–753.

Casenhiser, D.M.& A.E. Goldberg. 2005. Fast mapping of a phrasal form and meaning. *Developmental Science* 8. 500–508.

Church, K.W., W. Gale, P. Hanks, & D. Hindle. 1991. Using statistics in lexical analysis. In U. Zernik (ed), *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115–164. Hillsdale, NJ: Lawrence Erlbaum.

Colleman, T. & S. Bernolet. 2012. Alternation biases in corpora vs. picture description experiments: DO-biased and PD-biased verbs in the Dutch dative alternation. In D.S. Divjak & St.Th. Gries (eds.), *Frequency effects in language representation*, 87–126. Berlin & New York: De Gruyter Mouton.

*Collins Cobuild Dictionary of Idioms*. 2002. 2nd edition. London: Harper Collins.

Croft, W. 1998. Linguistic evidence and mental representations. *Cognitive Linguistics* 9. 151–173.

Croft, W. 2009a. Connecting frames and constructions: a case study of *eat* and *feed. Constructions and Frames* 1. 7–28.

Croft, W. 2009b. Toward a social cognitive linguistics. In V. Evans & S. Pourcel (eds.), *New directions in cognitive linguistics*, 395–420. Amsterdam & Philadelphia: John Benjamins.

Cruse, D.A. 1986. *Lexical semantics*. Cambridge: C.U.P.

Dąbrowska, E. 2009. Words as constructions. In V. Evans & S. Pourcel (eds.), *New directions in cognitive linguistics*, 201–223. Amsterdam & Philadelphia: John Benjamins.

Deese, J. 1964. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior* 3. 347–357.

Deshors, S.C. & St.Th. Gries. to appear. A case for the multifactorial assessment of learner language: the uses of *may* and *can* in French-English interlanguage. In D. Glynn & J. Robinson (eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy*. Amsterdam & Philadelphia: John Benjamins.

Divjak, Dagmar S. & St.Th. Gries. 2006. Ways of trying in Russian: clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2. 23–60.

Divjak, Dagmar S. & St.Th. Gries. 2008. Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon* 3. 188–213.

Divjak, Dagmar S. & St.Th. Gries. 2009. Corpus-based cognitive semantics: A contrastive study of phasal verbs in English and Russian. In K. Dziwirek & B. Lewandowska-Tomaszczyk (eds.), *Studies in cognitive corpus linguistics*, 273–296. Frankfurt am Main: Peter Lang.

Doğruöz, A.S. & St.Th. Gries. 2012. Spread of on-going changes in an immigrant language: Turkish in the Netherlands. *Review of Cognitive Linguistics* 10. 401–426.

Ellis, N.C. & R. Simpson-Vlach. 2005. An academic formulas list (AFL): extraction, validation, prioritization. Paper presented at Phraseology 2005, Université Catholique Louvain-la-Neuve.

Ellis, N.C. 2002. Frequency effects in language processing and acquisition. *Studies in Second Language Acquisition* 24. 143–188.

Ellis, N.C., R. Simpson-Vlach, & C. 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, N.C. & F. Ferreira-Junior. 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220.

Ellis, N.C. 2007. Language acquisition as rational contingency learning. *Applied Linguistics* 27. 1–24.

Ervin-Tripp, S. 1970. Substitution, context, and association. In L. Postman & G. Keppel (eds.), *Norms of word association*, 383–467. New York: Academic Press.

Fidelholtz, J.L. (1975), Word frequency and vowel reduction in English, *Chicago Linguistic Society* 11. 200–213.

Forster, K.I. & S.M. Chambers. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12. 627–635.

Fowlkes, E.B., & C.L. Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78. 553–569.

Frank, A.F., C. Kidd, M. Post, B. Van Durme, T.F. Jaeger. 2008. The web as a psycholinguistic resource. 5th International Workshop on Language Production. Annapolis, MD.

Gibbs, R.W. Jr. & Teenie Matlock. 2001. Psycholinguistic perspectives on polysemy. In H. Cuyckens & B. Zawada (eds.), *Polysemy in cognitive linguistics*, 213–239. Amsterdam & Philadelphia: John Benjamins.

Gilquin G. 2003. Causative 'get' and 'have'. So close, so different. *Journal of English Linguistics* 31. 125–148.

Glynn, D. 2010. Testing the hypothesis. Objectivity and verification in usage-based cognitive semantics. In D. Glynn & K. Fischer (eds.), *Corpus-driven cognitive semantics. Quantitative approaches*, 239–270. Berlin & New York: Mouton de Gruyter.

Goldberg, A.E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7. 219–224.

Goldberg, A.E. 2006. Constructions at work: On the nature of generalization in language. Oxford: O.U.P.

Goldberg, A.E. 1995. Constructions: A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Goldberg, A.E., D.M. Casenhiser, & N. Sethuraman. 2005. The role of prediction in construction learning. *Journal of Child Language* 32. 407–426.

Gries, St.Th. 2003. *Multifactorial analysis in corpus linguistics: a study of particle placement*. London: Continuum.

Gries, St.Th. under revision. Some current quantitative problems in corpus linguistics and a sketch of some solutions.

Gries, St.Th. & A. Stefanowitsch. 2004. Co-varying collexemes in the into-causative. In M. Achard & S. Kemmer (eds.), *Language, culture, and mind*, 225–236. Stanford, CA: CSLI.

Gries, St.Th. & A. Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9. 97–129.

Gries, St.Th. & A. Stefanowitsch. 2010. Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 73–90. Stanford, CA: CSLI.

Gries, St.Th. & C.V. David. 2007. This is kind of/sort of interesting: variation in hedging in English. In P. Pahta, I. Taavitsainen, T. Nevalainen, & J. Tyrkkoö (eds.). *Towards multimedia in corpus linguistics. Studies in variation, contacts and change in English 2*, University of Helsinki.

Gries, St.Th. & D.S. Divjak. 2010. Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal. In D. Glynn & K. Fischer (eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches*, 333–354. Berlin & New York: Mouton de Gruyter.

Gries, St.Th. & J. Mukherjee. 2010. Lexical gravity across varieties of English: an ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15. 520–548.

Gries, St.Th. & M. Hilpert. 2008. The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora* 3. 59–81.

Gries, St.Th. & M. Hilpert. 2010. Modeling diachronic change in the third person singular: a multi-factorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14. 293–320.

Gries, St.Th. & M. Hilpert. 2012. Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics. In T. Nevalainen & E.C. Traugott (eds.), The Oxford Handbook on the history of English, 134–144. Oxford: O.U.P.

Gries, St.Th. & N. Otani. 2010. Behavioral profiles: a corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34. 121–150.

Gries, St.Th. & S. Wulff. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3. 182–200.

Gries, St.Th. & S. Wulff. 2009. Psycholinguistic and corpus linguistic evidence for L2 constructions. *Annual Review of Cognitive Linguistics* 7. 163–186.

Gries, St.Th. 2003. Testing the sub-test: a collocational-overlap analysis of English *-ic* and *-ical* adjectives. *International Journal of Corpus Linguistics* 8. 31–61.

Gries, St.Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1–27.

Gries, St.Th. 2004. Isn't that fantabulous? How similarity motivates intentional morphological blends in English. In M. Achard & S. Kemmer (eds.), *Language, culture, and mind*, 415–428. Stanford, CA: CSLI.

Gries, St.Th. 2004. Shouldn't it be *breakfunch*? A quantitative analysis of the structure of blends. *Linguistics* 42. 639–667.

Gries, St.Th. 2004. Some characteristics of English morphological blends. In M.A. Andronis, E. Debenport, A. Pycha, & K. Yoshimura (eds.), *Papers from the 38th Regional Meeting of the Chicago Linguistics Society: Vol. II. The Panels*, 201–216. Chicago, IL: Chicago Linguistics Society.

Gries, St.Th. 2006. Cognitive determinants of subtractive word-formation processes: a corpus-based perspective. *Cognitive Linguistics* 17(4). 535–558.

Gries, St.Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In St.Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 57–99. Berlin & New York: Mouton de Gruyter.

Gries, St.Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1. 109–151.

Gries, St.Th. 2006. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54. 191–202.

Gries, St.Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13. 403–437.

Gries, St.Th. 2010. Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5. 323–346.

Gries, St.Th. 2010. Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, St.Th. 2010. Dispersions and adjusted frequencies in corpora: further explorations. In St.Th. Gries, Stefanie Wulff, & Mark Davies (eds.), *Corpus linguistic applications: current studies, new directions*, 197–212. Amsterdam: Rodopi.

Gries, St.Th. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In M. Brdar, St.Th. Gries, & Milena Žic Fuchs (eds.), *Cognitive linguistics: convergence and expansion*, 237–256. Amsterdam & Philadelphia: John Benjamins.

Gries, St.Th. 2012. Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language* 36. 477–510.

Gries, St.Th. 2012. Quantitative corpus data on blend formation: psycho- and cognitive-linguistic perspectives. In V. Renner, F. Maniez, & P. Arnaud (eds.), *Cross-disciplinary perspectives on lexical blending*, 145–167. Berlin & New York: Mouton de Gruyter.

Gries, St.Th. 2013. 50-something years of work on collocations: what is or should be next … *International Journal of Corpus Linguistics* 18. 137–165.

Gries, St.Th. 2013. *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin & New York: De Gruyter Mouton, p. 359.

Gries, St.Th., B. Hampe, & D. Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16. 635–676.

Gries, St.Th., B. Hampe, & D. Schönefeld. 2010. Converging evidence II: more on the association of verbs and constructions. In S. Rice & J. Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: CSLI.

Gries, St.Th., J. Newman, & Cyrus Shaoul. 2011. *N*-grams and the clustering of registers. *Empirical Language Research* 5.

Halliday, M.A.K. 2005. *Computational and quantitative studies*. London & New York: Continuum.

Hanks, P. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics* 1 75–98.

Hare, M.L., K. McRae, & J.L. Elman. 2003. Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language* 48. 281–303.

Harris, Z.S. 1970. *Papers in structural and transformational linguistics*. Dordrecht: Reidel.

Heeringa, W. 2004. Measuring dialect pronunciation differences using Levenshtein distance. Unpublished Ph.D. dissertation, University of Groningen.

Hilpert, M. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2. 243–57.

Hilpert, M. 2008. *Germanic future constructions. A usage-based approach to language change*. Amsterdam & Philadelphia: John Benjamins.

Hilpert, M. & St.Th. Gries. 2009. Assessing frequency changes in multi-stage diachronic corpora: applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 34. 385–401.

Hoey, M. 2005. *Lexical priming: a new theory of words and language*. London: Routledge.

Hommerberg, C. & G. Tottie. 2007. *Try to* and *try and*? Verb complementation in British and American English. *ICAME Journal* 31. 45–64.

Howes, D.H. & R.L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology* 41. 401–410.

Hunston, S. & G. Francis. 2000. *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam &Philadelphia: John Benjamins.

Janda, L., T. Nesset, & R.H. Baayen. 2010. Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling. *Corpus Linguistics and Linguistic Theory* 6. 29–48.

Kelly, M.H. 1998. *To brunch* or *to brench*: Some aspects of blend structure. *Linguistics* 36. 579–590.

Kishner, J.M. & R.W. Jr. Gibbs. 1996. How *just* gets its meanings: Polysemy and context in psychological semantics. *Language and Speech* 39. 19–36.

Kita, K., Y. Kato, T. Omoto, & Y. Yano. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing* 1. 21–33.

Krug, M. 1998. String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics* 26. 286–320.

Kubozono, H. 1990. Phonological constraints on blending in English as a case for phonology-morphology interface. Yearbook of Morphology 3. 1–20.

Langacker, R.W. 1987. *Foundations of Cognitive Grammar Vol. I: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, R.W. 1997. Constituency, dependency, and conceptual grouping. *Cognitive Linguistics* 8. 1–32.

Laufer, B. & T. Waldman. 2011. Verb-noun collocations in second language writing: a corpus analysis of learners' English. *Language Learning* 61. 647–672.

Leech, G.N. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82*, 105–122. Berlin & New York: Mouton de Gruyter.

Leech, G.N., P. Rayson, & A. Wilson. 2001. *Word frequencies in written and spoken English: based on the British National Corpus*. Longman: London.

Levelt, W.J.M. 1989. *Speaking: from thinking to articulation*. Cambridge, MA: The MIT Press.

Левенштейн, В. И. [Levenshtein, V.I.]. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10: 707–710.

López Rúa, P. 2002. On the structure of acronyms and neighbouring categories: A prototype-based account. *English Language and Linguistics* 6. 31–60.

Lyne, A.A. 1985. *The vocabulary of French business correspondence*. Geneva & Paris: Slatkine-Champion.

MacKay, D.G. 1973. Complexity in output systems: Evidence from behavioral hybrids. *American Journal of Psychology* 86. 785–806.

Malkiel, Y. 1959. Studies in irreversible binomials. *Lingua* 8. 113–160.

Margerie, H. 2008. A historical and collexeme analysis of the development of the compromiser *fairly*. *Journal of Historical Pragmatics* 9. 288–314.

Marslen-Wilson, William. 1987. Functional parallelism in spoken word-recognition. *Cognition* 25. 71–102.

Mason, O. 2006. The automatic extraction of linguistic information from text corpora. Unpublished Ph.D. dissertation, University of Birmingham.

Mason, O. 2007. From lexis to syntax: The use of multi-word units in grammatical description. *Proceedings of Lexis and Grammar 2007, Bonifacio, Corsica*.

McClelland, J.L., & D.E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review* 88. 375–407.

McDonald, S., R. Shillcock, & C. Brew. 2001. Low-level predictive inference in reading: Using distributional statistics to predict eye movements. Paper presented at the 7th Annual Conference on Architectures and Mechanisms for Language Processing, Saarbrücken.

McDonald, S. 1997. Exploring the validity of corpus-derived measures of semantic similarity. Paper presented at the 9th Annual CCS/HCRC postgraduate conference, University of Edinburgh.

McEnery, T. & A. Wilson. 1996. *Corpus linguistics*. Edinburgh: E.U.P.

McEnery, T., R. Xiao, & Y. Tono. 2006. *Corpus-based language studies: an advanced resource book*. London & New York: Routledge.

Meyer, C.F. 2002. *English corpus linguistics: an introduction*. Cambridge: C.U.P.

Miller, G.A. & W.G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6. 1–28.

Mintz, T.H., E.L. Newport, & T.G. Bever. 2002. The distributional structure of grammatical categories in the speech to young children. *Cognitive Science* 26. 393–424.

Mollin, S. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5. 175–200.

Norvig, P. & G. Lakoff. 1987. Taking: A study in lexical network theory. *Berkeley Linguistics Society 13*, 195–206.

Noteboom, S.G. 1981. Lexical retrieval from fragments of spoken words: Beginnings vs. endings. *Journal of Phonetics* 9. 407–224.

Onnis, L., P. Monaghan, K. Richmond, & N. Chater. 2005. Phonology impacts segmentation in online processing. *Journal of Memory and Language* 53. 225–237.

Partington, A. 1998. *Patterns and meanings: using corpora for English language research and teaching*. Amsterdam & Philadelphia: John Benjamins.

Pierrehumbert, J. 2003. Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay & S. Jannedy (eds.), *Probabilistic linguistics*, 177–228. Cambridge, MA: The MIT Press.

Pound, Louise. 1914. *Blends: Their relation to English word-formation*. Heidelberg: Winter.

Raukko, J. 1999. An intersubjective method for cognitive-semantic research on polysemy: the case of *get*. In: M.K. Hiraga, C. Sinha, & S. Wilcox (eds.), *Cultural, psychological and typological issues in cognitive linguistics*, 87–105. Amsterdam & Philadelphia: John Benjamins.

Raukko, J. 2003. Polysemy as flexible meaning: experiments with English *get* and Finnish *pitää*. In B. Nerlich, Z. Todd, V. Herman, & D.D. Clarke (eds.), *Polysemy: flexible patterns of meaning in mind and language*, 161–193. Berlin & New York: Mouton de Gruyter.

Raymond, W.D. & E.L. Brown. 2012. Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In St.Th. Gries & D.S. Divjak (eds), *Frequency effects in language learning and processing*, 35–52. Berlin & New York: De Gruyter Mouton.

Reddington, M., N. Chater, & S. Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science* 22. 435–469.

Rice, Sally. 1996. Prepositional prototypes. In M. Pütz & R. Dirven (eds.), *The construal of space in language and thought*, 135–165. Berlin & New York: Mouton de Gruyter.

Roland, D. & D. Jurafsky. 2002. Verb sense and verb subcategorization probabilities. In S. Stevenson, & P. Merlo (eds.), *The lexical basis of sentence processing: formal, computational, and experimental issues*, 325–346. Amsterdam & Philadelphia: John Benjamins.

Roland, D., F. Dick, & J.L. Elman. 2007. Frequency of basic English grammatical structures: a corpus analysis. *Journal of Memory and Language* 57. 348–379.

Saffran, J.R., R.N. Aslin, & E.L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274 (5294), 1926–1928.

Sandra, D. & S. Rice. 1995. Network analyses of prepositional meaning: mirroring whose mind—the linguist's or the language user's? *Cognitive Linguistics* 6. 89–130.

Schmid, H.-J. 2000. *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin & New York: Mouton de Gruyter.

Schmid, H.-J. 1993. Cottage and co., idea, start vs. begin. Tübingen: Max Niemeyer.

Schuchardt, H. 1885. *Über die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Robert Oppenheim.

Simpson, R. & N.C. Ellis. 2005. An academic formulas list: Extraction, validation, prioritization. Paper presented at Phraseology 2005, Université Catholique de Louvain.

Sinclair, J.M. 1991. *Corpus, concordance, collocation*. Oxford: O.U.P.

Snider, N. 2009. Similarity and structural priming. *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, 815–820.

Stefanowitsch, A. 2011. Cognitive linguistics as a cognitive science. In M. Callies, W.R. Keller, & A. Lohöfer (eds.), *Bi-directionality in the cognitive sciences: avenues, challenges, and limitations*, 295–310. Amsterdam & Philadelphia: John Benjamins.

Stefanowitsch, A. & St.Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8. 209–243.

Stefanowitsch, A. & St.Th. Gries. 2003. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8. 209–243.

Stoll, Sabine & St.Th. Gries. 2009. How to measure development in corpora: an association strength approach. *Journal of Child Language* 36. 1075–1090.

Stubbs, M. 1993. British traditions in text analysis: From Firth to Sinclair. In M. Baker, F. Francis & E. Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 1–46. Amsterdam & Philadelphia: John Benjamins.

Szmrecsanyi, B. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 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 & New York: Mouton de Gruyter.

Teubert, W. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10. 1–13.

Theijssen, D., L. ten Bosch, L. Boves, B. Cranen, & H. van Halteren. 2013. Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9. 227–262.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam & Philadelphia: John Benjamins.

Tomasello, Michael. 2005. *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: H.U.P.

Tryk, H.E. 1986. Subjective scaling of word frequency. *The American Journal of Psychology* 81. 170–177.

Tyler, A. & V. Evans. 2001. Reconsidering prepositional polysemy networks: The case of *over*. Language 77. 724–765.

Wiechmann, D. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4. 253–290.

Wilcox, Rand. 2012. *Modern statistics fr the social and behavioral sciences: a practical introduction*. Boca Raton, FL: CRC Press.

Williams, G. 2006. La linguistique de corpus: Une affaire préposition—nelle. In F. Rastier & M. Ballabriga (eds.), *Corpus en Lettres et Sciences Sociales: Des Documents Numériques à l'Interprétation*, 151–158. Paris: Texto.

Wulff, S., A. Stefanowitsch, & St.Th. Gries. 2007. Brutal Brits and persuasive Americans: variety-specific meaning construction in the *into*-causative. In G. Radden, K.-M. Köpcke, T. Berg, & P. Siemund (eds.), *Aspects of meaning construction*, 265–281. Amsterdam & Philadelphia: John Benjamins.

Wulff, S. 2006. *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In St.Th. Gries & Anatol Stefanowitsch (eds.). *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 101–125. Berlin & New York: Mouton de Gruyter.

Xiao, R. 2009. Theory-driven corpus research: Using corpora to inform aspect theory. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, 987–1008. Berlin & New York: Mouton de Gruyter.

Yuen, K.K. 1974. The two sample trimmed t for unequal population variances. *Biometrika* 61. 165–170.

# Important Resources 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), CIFCL is one of the most important international events in the field of Cognitive Linguistics. It is supported by international Cognitive Linguistics community and attended by a large number of researchers. Organizer: Thomas Li thomasli@buaa.edu.cn Book Series: Eminent Linguists Lecture Series (with DVD)

4. http://www.novapublishers.com/catalog/product_info.php?products_id=10019
Website for the *International Journal of Cognitive Linguistics*, edited by CIFCL

5. http://www.degruyter.com/view/serial/16078?rskey=fw6Q2O&result=1&q=CLR
Website for the Cognitive Linguistics Research [CLR], edited by Dirk Geeraerts and John R. Taylor. Honorary editors: René Dirven, Ronald W. Langacker

6. http://www.degruyter.com/view/serial/20568?rskey=dddL3r&result=1&q=ACL
Website for Application of Cognitive Linguistics [ACL], edited by Gitte Kristiansen and Francisco J. Ruiz de Mendoza Ibáñez

7. http://www.benjamins.com/#catalog/books/clscc/main
Website for book series in Cognitive Linguistics Studies in Cultural Contexts

8. http://www.degruyter.com/view/db/cogbib
Website for online resources for Cognitive Linguistics Bibliography

9. http://benjamins.com/online/met/
Website for Bibliography of Metaphor and Metonymy

10. http://linguistics.berkeley.edu/research/cognitive/
Website for Language and Cognition in Berkeley

11. https://framenet.icsi.berkeley.edu/fndrupal/
    Website for FrameNet

12. Founding Fathers of Cognitive Linguistics

    Leonard Talmy
    http://linguistics.buffalo.edu/people/faculty/talmy/talmyweb/index.html

    Ronald W. Langacker
    http://idiom.ucsd.edu/~rwl/

    George Lakoff
    http://georgelakoff.com/