Draft of January 1, 2014. Under review for Nancy Ide and James Pustejovsky, eds., *The Handbook of Linguistic Annotation*. Comments and corrections welcome.

Developing linguistic theories using annotated corpora*

Marie-Catherine de Marneffe and Christopher Potts

Abstract This paper aims to carve out a place for corpus research within theoretical linguistics and psycholinguistics. We argue that annotated corpora naturally complement native speaker intuitions and controlled psycholinguistic methods and thus can be powerful tools for developing and evaluating linguistic theories. We also review basic methods and best practices for moving from corpus annotations to hypothesis formation and testing, offering practical advice and technical guidance to researchers wishing to incorporate corpus methods into their work.

1 Introduction

Annotated corpora can be powerful tools for developing and evaluating linguistic theories. By providing large samples of naturalistic data, such resources complement native speaker intuitions and controlled psycholinguistic methods, thereby putting linguistic hypotheses on a sturdier empirical foundation. Corpus data and methods also open up new analytic and methodological possibilities, which can broaden the scope of linguistics and increase its relevance to language technologies and neighboring scientific fields.

With this paper, we aim to carve out a place for corpus research within theoretical linguistics and psycholinguistics. We have the impression that, within these communities, annotated corpora are often regarded as irrelevant — useful for building computational models and exploring theories of corpus linguistics, but unhelpful when it comes to pursuing questions about language structure and language processing.

Marie-Catherine de Marneffe
Department of Linguistics, The Ohio State University, e-mail: mcdm@ling.ohio-state.edu

Christopher Potts
Department of Linguistics, Stanford University, e-mail: cgpotts@stanford.edu

* Thanks to David Beaver, Nancy Ide, Dan Lassiter, James Pustejovsky, ...

The disciplinary boundaries are sometimes even more firmly drawn, with corpus research portrayed as incompatible with foundational assumptions about linguistic inquiry, fundamentally limited in the kinds of evidence it can provide, and at odds with established methods for conducting psychological experiments. Of course, many linguists have embraced corpus work, but negative perceptions remain prominent.

In secs. 2–3, we address these concerns, arguing that they are misplaced and showing that corpora can be used to inform a wide range of hypotheses. We also seek to provide practical advice and technical guidance to linguists wishing to incorporate corpus methods into their work. To this end, sec. 4 reviews different sources for annotations and different kinds of annotation project, and sec. 5 outlines basic methods and best practices for moving from corpus data to hypothesis formation and testing. Throughout these discussions, we emphasize that all annotations are the product of theoretical assumptions, complex social factors, and linguistic intuitions, and we argue that these interacting factors should be identified and clearly reflected in how hypotheses are formulated and assessed.

This paper is intended as a companion to others in this volume, which review specific annotation schemes and corpora. Our focus is on the conceptual issues surrounding using corpora for linguistic work: finding the right kinds of annotated data, navigating large and unruly corpora, moving from intuitive general hypotheses to corpus-specific ones, and relating corpus results to theoretical ideas. Thus, we do not discuss specific corpora, annotation schemes, or projects in any detail. Our aim is rather to motivate a general analytic framework, and to highlight papers that use corpora in diverse ways to tackle subtle theoretical questions.

This is an opportune moment. To some extent, corpus investigations have already made their way into theoretical linguistics, as linguists search the Web with the goal of showing that theoretically informative phenomena are attested. While this has had a profound, positive effect on linguistics, it has strained the field's relationship with current search engines. Industrial search engines deal primary in byte streams (or, at best, whitespace-delimited lists of characters). Linguists know better than anyone that these objects are mere blurry reflections of the conceptual units of natural language: phones, words, phrases, sentences, utterances, and so forth. The Web-searching linguist is liable to grow dissatisfied quickly. With the present paper, we hope to capitalize on this dissatisfaction, by pointing the way to richer corpus investigations involving annotated data and a fuller stock of investigative methods.

2 Corpus investigations in the context of linguistic theory

This section seeks to situate corpus work within the broader context of linguistic theory and related fields. Our goal is to show that corpus investigations, considered as complex measurements, observational studies, or natural experiments, are compatible with a wide variety of approaches to theorizing about language. We also critically assess claims about the methodological limitations of corpus research.

2.1 Intuition and experiment

Experimental and corpus methods are often defined in opposition to 'intuition-based' (introspective, armchair) methods. We think this framing of the issues is misleading. All scientific inquiry is driven by the investigators' intuitions about the world. In linguistics, these intuitions are often those of a native-speaker scientist or her trusted consultants, and such intuitions are probably rightly privileged for their nuance, depth, and accuracy (for discussion, see Schütze 1996; Devitt 2006; Culbertson & Gross 2009). All successful corpus investigations are guided by such intuitions, which shape the annotations and guide their analysis. The same can be said of psycholinguistic experiments, where native speaker intuitions shape the experimental items and the interpretation of results.

There is an aspect of the intuition vs. experimentation framing that we do endorse: introspection should be the *start* of the investigation, not its culmination. Like any source of evidence, intuitions are fallible. Their limitations become especially apparent where theoretical goals and preferences are part of the picture (Spencer, 1973; Gordon & Hendrick, 1997). Corpus research can serve as an important check against such biases, by bringing in large quantities of data that were not produced by the investigators. More generally, intuitions should be followed by further and more systematic inquiry, using corpora or psycholinguistic experiments — preferably both!

2.2 Corpora and experimental methods

One way to address the question of how corpus research relates to psycholinguistics is to consider whether corpora can support experiments that conform to the norms and best-practices of psychology, a parent field of psycholinguistics.

Winston & Blais (1996) study how the concept of an experiment is identified in textbooks in the period 1930–1970, in psychology, sociology, biology, and physics. They see three general kinds of definition recurring in these texts (p. 603–604):

1. An empirical or systematic study, or data collection, with no mention of control or manipulation.
2. Observations or repeated observations under controlled or standardized conditions, with no mention of manipulation.
3. Manipulation of a factor or variable while controlling or holding all others constant.

According to Winston & Blais's survey, by the 1970s, definition 3 was more or less fully established in psychology textbooks (and, to a lesser extent, sociology textbooks), with many texts explicitly contrasting it with notions like observation, correlation, and introspection. However, throughout the same time period, biology

and physics remained dominated by more general definitions like 1 and 2.[2] In particular, Winston & Blais (1996:606) say, "Physics texts often describe the precise measurement of a quality or measurement to test a theoretical prediction as examples of experiment". This is close to the notion of experiment that is likely to be in play when one works with corpora, since the corpus researcher rarely has the chance to do the sort of active manipulation that is central to definition 3.

Two qualifications are in order, though. First, crowdsourcing (Poesio, this volume) has made it possible to annotate vast amounts of data relatively quickly and inexpensively, paving the way for annotation projects to use psycholinguistic methods in both the design and analysis phases. The differences between such projects and a standard human-subjects experiment might lie entirely in the kinds of data used — hand-crafted examples in the case of experiments and naturalistic data in the case of annotation projects. For example, de Marneffe et al. (2010), de Marneffe et al. (2012), and Degen (2013) crowdsourced dozens of annotations for each of their corpus examples and used the annotation/response distributions to characterize and predict communicative uncertainty. Similarly, Potts (2012b) essentially uses a between-subjects design to record, in a metadata-rich corpus, the effects of different contextual constraints on crowdsourced workers' interactions, a paradigm case of the kind of active manipulation that characterizes definition 3. (For additional discussion, see sec. 4.5.)

Second, in fields like sociology, political science, and economics, definition 3 will often be unobtainable for the same reasons that it is unobtainable in corpus research: the object of study is a set of past events, and reproductions of those events in the lab are either impossible or impractical. Here, the very nature of the inquiry forces the studies to be observational. Causal inference is often still a goal in such situations, so statistical models have been developed that support causal inferences even in the absence of pre-defined, randomly selected control and treatment groups, uniform experimental settings, and active manipulation. See Gelman 2011 for a review of the issues and current approaches to causal inference in both experimental and observational contexts.

At any rate, definition 3 is a special case of the other two, imposing more stringent requirements and typically licensing stronger inferences. Whether a corpus study can or must rise to this level seems best addressed on a case-by-case basis, in the context of what the research questions are like and what data are available.

2.3 What is a corpus?

Our emphasis is on the role that corpora can play in developing linguistic hypotheses, so it behooves us to be permissive in specifying what counts as a corpus. Thus,

[2] Winston & Blais suggest that the underlying causes of these differences are complex, relating to the practices of sub-disciplines within these fields, the role of causal inference in building theories, and perceived needs to be rigorous (biology and physics textbooks and lab manuals are much more likely not to address these methodological questions at all).

we say that a corpus is any collection of language data (Kilgarriff & Grefenstette, 2003). We leave open the origin of this data, its size, its basic units, and the nature of the data that it encodes, which could come in any medium. We even count as corpora things like dictionaries, specialized word lists (Dewey, 1923; Zipf, 1949; Wierzbicka, 1987; Levin, 1993; Hoeksema, 1997; Michel et al., 2011), and aggregated linguistic judgments (Sprouse et al., 2013), which do not represent specific linguistic events, but rather aim to encode general features of the linguistic system. More specialized definitions would only limit the kinds of questions one can address, which runs against our goals in this paper. We are similarly open about what counts as an annotation (sec. 4).

Though we do not adopt restrictive definitions, we are extremely concerned with the ways in which the properties of specific corpora relate to the kinds of questions one can address with them and the strength and persuasiveness of the resulting claims. From this perspective, it makes sense to try to work with corpora of the sort defined by Gries and Berez, this volume, and McEnery & Wilson 2001:§2: balanced, representative of the population under investigation, and produced under conditions that align with the empirical goals of the study.[3] These are ideals, though; since we lack robust criteria for deciding whether a corpus manifests them (Kilgarriff & Grefenstette, 2003), the most productive thing one can do is report the properties of one's corpus as comprehensively as possible. (See sec. 5.2 for related discussion.)

2.4 Conceptual foundations

Within linguistics and the philosophy of language, there is continued debate about the nature of the objects under investigation. Are they events in the world, events in the brain, abstract objects, or community-wide conventions? There is not space for us to seriously engage this issue (see Jackendoff 1992; Harris 1993; Lassiter 2008; Scholz et al. 2011), but it is worth raising here, because corpus methods are sometimes unfairly branded as involving commitments about this foundational question. In fact, corpus methods are compatible with all of the major positions on this issue.

The *nominalist* position is that linguists should study tokens: linguistic events in the world as encoded in texts, sound recordings, and so forth. This position can arise either from ontological skepticism about abstract objects or methodological skepticism about our ability to achieve a scientific understanding of abstract objects. In linguistics, nominalism is closely related to strongly behaviorist stances in psychology, which hold that we can objectively study only observable behavior. Purely nominalist theories of language like that of Harris (1954) hold that all theoretical claims must take the form of statements about distributions of tokens in sample data; extrapolations from the tokens to types (phonemes, words, etc.) are meant to have no theoretical status.

[3] In general, one hopes that the speakers who contributed to the corpus were unconstrained by non-linguistic factors like editorial rules, censorship, and other performance limitations, but we can imagine studies where such factors actually serve the investigative goals.

Externalists take exactly the opposite view: they embrace abstractions from the tokens we encounter to abstract objects like types; the tokens themselves likely have no theoretical standing, serving only as a means for discovering the abstractions. Within externalism, *conventionalist* views regard language as a system of conventions, inhering in no individual's head, but rather existing only at the community-level (Lewis, 1969; Putnam, 1975; Burge, 1979), whereas the *platonist* view is that linguistic objects are abstract mathematical objects that individuals can have knowledge of (Katz, 1981; Katz & Postal, 1991).

Chomsky (1957a,b, 1986) famously rejected all of these views, seeking to replace them with an *internalist* or *mentalist* position in which linguistics is the science of individuals' mental capacity to learn and process language. From this perspective, it is useful to study linguistic objects and their use only insofar as such study yields insights into speakers' cognitive abilities. As with nominalism, the abstract linguistic objects have no status in the theory, though not out of skepticism that such abstract objects exist but rather out of a belief that they are irrelevant to the science of linguistics. Similarly, community-wide conventions play no role in the theory; they shape individuals' linguistic abilities, but they are not the object of study.

While advancing his internalist position, Chomsky targeted corpus methods, associating them with nominalism and externalism. This connection might seem warranted at first: for the most part, corpora consist of partial recordings of specific linguistic events involving numerous individuals, so corpus results might seem doomed to be results about tokens or populations. However, we reject this conflation. Corpus research is compatible with all of the above theoretical perspectives, and thus doing corpus research brings with it no commitments on this point.

What the nominalist classification of corpus work misses is the role of inference and generalization. Where the corpus is the ultimate object of study, the theoretical stance is likely to be nominalist. However, according to McEnery & Wilson (2001:7), even early corpus linguists sought to use corpora primarily to formulate predictions about new data (Hockett, 1948, 1954). In modern work, the corpus is essentially never the primary object of study, but rather only a source of evidence for more general claims. Those claims can be made in terms of abstract objects, mental constructs, or conventions (perhaps among other possibilities).

What the externalist classification misses is the freedom one has in choosing or collecting one's corpora. For the most part, corpora consist of data from a variety of speakers, so generalizations extracted from them will most easily be phrased as generalizations about populations, a natural fit for conventionalism. However, there are also corpora that represent single individuals — a person's diary, an author's collected works, the set of all email messages sent by an individual in a given year, and so forth. Corpora can be extremely broad or incredibly fine-grained; as with other modes of inquiry, we are limited only by our ability to gather evidence, and the nature of the evidence we collect will constrain the kinds of inferences we can make with confidence.

We are not surprised that the Chomsky of 1957 regarded corpus research as anathema to his internalist, mentalist program. In its current form, corpus research is heavily dependent on information theory (Cover & Thomas, 1991), which was

only in early development itself in the 1950s (Shannon, 1948). So, in 1957, corpus research probably did look mainly like a lot of counting for its own sake. However, the situation is radically different now. Corpus research is every bit as theory-driven as theoretical linguistics, and it has strong and well-understood mathematical foundations. It is thus surprising to find that Chomsky is as strident as ever about corpus research, saying, for example that it "doesn't mean anything" and characterizing it as just an attempt to "accumulate huge masses of unanalyzed data and to try to draw some generalization from them" (Andor 2004:97). On the positive side, though, he does say, "We'll judge it by the results that come out" (p. 97). This is the view we advocate for all approaches to gathering evidence, and we think corpus methods will fare well in this judgment.

2.5 Competence and performance

Chomsky and others have also criticized corpus methods for being unable to distinguish competence (the abstract cognitive ability speakers have) from performance (the regular use of language). The rationale behind this criticism seems to be as follows: corpora are records of specific instances of language use, and thus they will inevitably contain distracting phenomena and patterns that derive entirely from issues of performance — for example, speech errors and disfluencies, frequency distributions derived from real-world goals rather than linguistic pressures, and short-term memory limitations (Chomsky 1986; Leech 1992; McEnery & Wilson 2001:6).

Our response here (as with so many of these foundational issues) is that corpus methods are not specially problematic for linguists wishing to distinguish competence from performance. It is well-known, for example, that speakers' introspective judgments will be shaped by non-linguistic factors, including cognitive load, the social dynamics of the situation, fatigue, inebriation, and repeated exposure (Snyder, 2000). These same worries pertain to laboratory situations, in which subjects can suffer from all of these cognitive limitations, and the experimenters themselves might inadvertently introduce factors into the experimental situation that get in the way of observing competence. In all these cases, the only antidote is care — care with the materials, participants, and analysis. If we adopt the terms of the competence/performance distinction, then we must confront the fact that all our experience with language, whether introspective or interactive, is via performance data (Chomsky 1965:11, cited by Scholz et al. 2011).

The other side of this issue is that performance is important in its own right, not only for what it can tell us about language production (Jurafsky, 1996; Levy & Jaeger, 2007; Frank & Jaeger, 2008) and comprehension (Levy, 2008), but also for understanding the nature of competence itself (Sag & Wasow, 2011). Here, corpora have proven invaluable in part because they are likely to encode errors in ways that allow us to glimpse the systematic cognitive processes that contribute to them. This is perhaps nowhere more evident than in child language acquisition, where the CHILDES corpus (MacWhinney, 2000) has long been used to gain insights into

children's linguistic knowledge at various stages of development, often by observing their performance errors.

Errors are a source of insights for adult sentence processing as well. For example, subject–verb agreement errors from corpora have played a role in developing not only models of sentence processing but also formal models of morphosyntactic feature sharing (Bock et al., 2006; Frazier, 2012). Similarly, unintentionally overnegated structures (*no head injury is too trivial to ignore*; Wason & Reich 1979; Horn 1991; Barton & Sanford 1993) have long been a source of insights into the relationship between encoded content and intended content (Clark, 1997). Errors of comprehension can be equally enlightening. For instance, corpora of misheard song lyrics can inform theories of acoustic phonetics, auditory perception, and phonological feature structures (Vitevitch, 2002; Ring & Uitdenbogerd, 2009). The common theme of all these cases is that corpora often reveal systematicity in people's performance errors, which can provide a clear window into competence.

2.6 Statistical measures and scientific generalizations

For the most part, evidence gathered from corpora will have a statistical quality. We rarely observe categorical phenomena, but rather gradations. In probabilistic approaches (Jurafsky, 1996; Bod et al., 2003; Goodman & Lassiter, To appear), it might be possible to incorporate such non-categorical values directly into the theory or use them directly when assessing theoretical hypotheses. In non-probabilistic approaches, the status of intermediate probability values might be less evident, and this might lead one to infer that such values conflict with such approaches.

We argue that this inference would be incorrect; corpus work imposes *no* theoretical commitments on this point. On the one hand, one can view the statistical patterns as reflecting underlying stochastic processes. On the other hand, one might view them as reflecting the interaction of a diverse set of fundamentally categorical restrictions, perhaps further affected by issues that fall outside of the theory (Manning 2003:§3.1). From this perspective, if we could isolate all of the categorical restrictions and remove issues of performance, we would see categorical phenomena. Broadly speaking, this kind of position is not as unusual as one might think; even in thoroughly probabilistic theories like quantum mechanics, there is apparently still debate about whether the underlying principles have a stochastic component (Faye, 2008). Within linguistics, precisely these interpretive issues came to the fore in a recent, widely-observed debate about the nature of question formation and other long-distance dependencies: Hofmeister & Sag 2010; Hofmeister et al. 2012a,b; Sprouse et al. 2012a,b; Sprouse & Hornstein 2013.

2.7 From unattested to impossible

Corpus-based research is often criticized for being able to support conclusions only about what is possible, not what is impossible. There is a sense in which this is true, but it is unfair to single our corpus methods on this point. This limitation is shared by all empirical methodologies and approaches, which should come as no surprise, since it is just an instance of the limitations of scientific induction (Vickers, 2013). In the context of linguistic theory, we emphasize that intuitions too can be fallible; an analyst's judgment that something is impossible might be correct, or it might simply be a failure of imagination (Fillmore, 1992; Manning, 2003). Similarly, psycholinguistic methods cannot (and do not purport to) offer proof of impossibility. In all these cases, we must risk the step from a finite amount of evidence to a claim that something is ruled out in principle. For intuition-driven research, the evidence consists of a finite set of psychological reactions. For psycholinguistics, it consists of a finite set of reactions from subjects. For corpus research, it consists of the corpus data. Each kind of inference comes with its own limitations, risks, and advantages.

2.8 Corpus research and natural language processing

Many corpora (including most of those discussed in this volume) were developed primarily to train and evaluate computational models and implemented systems, as part of the field of natural language processing (NLP; Manning & Schütze 1999; Jurafsky & Martin 2009). Such research is often subtly different from linguistic research. Linguists typically formulate very specific hypotheses and try to evaluate them in focused ways, whereas NLP assessments tend to be holistic. The linguist might not care that her hypothesis is relevant to only a small part of the data, as long as it has no exceptions, whereas the NLP researcher typically aims to account for the whole of a particular data set and might not worry about a few exceptions. However, we do not want to make too much of the difference. All things considered, the NLP researcher would like her model to provide deep insights, and the linguist would like to give a comprehensive account. The differences we just mentioned are thus ones of emphasis and focus in daily practice.

The two modes of inquiry naturally complement each other as well. This is particularly true in the context of current statistical approaches to NLP, in which the models can include vast numbers of features and the training phase involves inferring, from the available data, which features matter and how they interact. Thus, the NLP researcher can often incorporate diverse theoretical ideas as part of her feature extraction function (see sec. 5.3), and the NLP evaluation serves as one kind of assessment of those ideas. The examples of this fruitful dynamic between NLP and theoretical linguistics are too numerous for us to enumerate. Suffice it to say that it has played an important role in the rapid progress in computational phonology (Goldwater & Johnson, 2003; Hayes & Wilson, 2008), morphological analysis (Goldsmith, 2001; Goldwater et al., 2006; Roark & Sproat, 2007; Munro, 2012),

semantic parsing (Wong & Mooney, 2007; Zettlemoyer, 2009; Kwiatkowski et al., 2011; Liang et al., 2013), and anaphora resolution and discourse coherence (Gordon et al., 1993; Walker et al., 1997; Beaver, 2004, 2007), among many other areas. Increasingly, linguists are incorporating probabilistic ideas into their theories, and NLP researchers are embracing highly structured representations, so we expect to see further cross-pollination between these two fields.

3 Hypothesis formation in the context of corpus work

This section addresses the question of what kinds of hypotheses one can pursue using corpora. The discussion is framed around three kinds of very general hypothesis: *X* is possible (grammatical, meaningful, felicitous), *X* is impossible (ungrammatical, meaningless, infelicitous), and *X* is (un)likely, (dis)preferred, or (un)marked.

3.1 Possible

As we noted in sec. 1, recent linguistic research has been shaped, for the better, by the growth of the Web and the existence of powerful search engines. The primary way in which the Web searches fuel such research is by turning up attested instances of certain phenomena. More generally, corpora excel at showing that certain things are possible, and it is now easy to point to cases where this has played a pivotal role in linguistic debates (Hoeksema, 2008; Potts, 2012a; Glass, 2013; Grimm & McNally, 2013). A few words of caution are in order here, however.

First, depending on the nature of the corpus, it might be crucial for native speakers to provide their judgments of the examples in question (Schütze, 2009). This is less pressing for highly structured, carefully collected corpora, but it is essential for messy, unstructured ones, for example, those derived from the Web. Native speaker judgments will combat problems relating to mis-interpreting the data, which can arise when one mistakes one phenomenon for another, treats an error as a genuine example, or misconstrues word-play and other non-literal uses.

The above assumes that it is possible to inspect all of the relevant examples, judging each one and making decisions accordingly. This is not always an option. The corpus might be too large for this to be practical; or it might only partially represent the underlying data, leaving crucial information out; or finding speakers of the relevant dialects might be hard. In such situations, it is more difficult to determine whether the examples one has found are truly systematic or represent mere idiosyncrasies in the data, which can arise from a host of irrelevant and partly random processes (encoding errors, typographic mistakes, performance errors, etc.). Any sufficiently large data set is bound to contain such errors (Rajaraman & Ullman 2011:§1).

A rich, well-defined theoretical model is the best defense against spurious conclusions about what's possible. Together with careful handling of the data (sec. 5), a model can quantify the strength of the evidence and thus lead to stronger evaluations. Manning (2003) provides a useful illustrative example. He reports being surprised upon reading *as least as* where he expected *at least as*. Does this represent a genuine point of variation, or is it a mere typo? Manning's subsequent searches with large corpora and the Web turned up hundreds of additional examples, suggesting that the form is genuine, but the denominator (the amount of text being searched) is growing as the stock of attested examples grows (Schütze, 2009). To more systematically explore the likelihood that these attested examples are genuine, we might compare the observed corpus frequencies with other factors — for example, our estimate of the probability of typing 's' where 't' was intended, and psycholinguistic evidence relating to conceptual mistakes (e.g., is the initial *as* a reflex of the speaker's planning for a later comparative like *as tall as*?).

The previous example addresses the question of how reliable a given set of tokens is. In the context of a statistical model, corpora can also be used to motivate claims in the other direction: that specific phenomena are possible even though they are not directly attested in the data. For instance, a model trained on data containing a subject–verb combination (S,V) and a verb–object combination (V,O) might predict that (S,V,O) is licit even though it never appears in the data (Pereira, 2000; Norvig, 2011). In this case, the corpus itself does not show that (S,V,O) is attested, but, together with the model, it makes a prediction about that form. If the prediction passes muster with native speakers and competent experimental participants, then we might feel confident in it (and perhaps feel increased confidence in our model).

3.2 Impossible

In sec. 2.7, we pointed out that, like all methods, corpus investigation can motivate inductive, not deductive, generalizations, and thus universal generalizations are always risky. Nonetheless, there are analytic steps one can take, in the context of corpus work, to mitigate this risk.

Perhaps the most important step is ensuring that the corpus is properly aligned with one's scientific hypothesis. If one is studying slang forms, the financial pages of major newspapers are unlikely to provide a good fit — the absence of a specific form could be explained by differences in register, social norms, etc. The better the fit between corpus and hypothesis, the less likely it is that the absence of a form has alternative explanations tracing to sampling errors.

As above, a specific model, together with a corpus, might support claims that something is impossible. A given form might be both absent from the data and predicted by the model to have vanishingly low probability compared with others. This might further license the step of calling the form impossible, especially if one can identify features of the data and model that lead to this prediction. Pereira (2000) uses such reasoning to argue that a simple statistical model, trained on newspaper

text, predicts *furiously sleep ideas green colorless* to be impossible, or at least dramatically less likely than *colorless green ideas sleep furiously*, thereby answering a challenge from Chomsky (1957b).

3.3 Biases, preferences, and markedness

Speakers display preferences and biases in production and construal, at all levels of linguistic description. Corpora are ideal for capturing such patterns and can be an important counterpart to preference data collected in the lab — the corpora record (perhaps messily) a wide range of contexts, interpersonal situations, and psychological constraints, while the experimental data represent highly controlled (perhaps artificial) scenarios. Information about preferences is often left out of linguistic theory, which excels at saying simply that multiple options are available, but corpus methods allow us to bring the relevant information into the model.

In sec. 2.6 above, we argued that corpus methods do not entail a probabilistic approach to linguistic theory. Nonetheless, working with corpora is likely to make one feel more receptive to probabilistic hypotheses. The issue is that any non-trivial claim one makes about language is likely to be falsified, in the categorical sense, given a sufficiently large corpus, even assuming rigorous criteria such as those reviewed in sec. 3.1. However, it is a great loss to simply say, in the face of a handful of examples out of millions, that the proposed hypothesis is false. It might capture a deep and important regularity, so we should be encouraging about finding a place for it in our theories.

Bringing probabilistic statements into linguistic theory does not need to be as dramatic a move as it sounds. In many cases, it is conceptually and theoretically natural to assume a division between the categorical and non-categorical components. In phonology, statistical regularities in the lexicon of a language can be construed as providing evidence for a probabilistic grammar, but they might also be seen as capturing information about markedness, a concept that can be modeled in non-probabilistic terms at the level of grammatical typologies and the path of language acquisition. In morphology and syntax, the grammar rules capture what is possible, and associated weights or probabilities capture their frequency of use in real data. For examples of morphosyntactic analyses that are compatible with such views, see Sproat & Shih 1991; Manning 2003; Bresnan & Nikitina 2010; Levy 2008; Thuilier et al. 2013. Similarly, in the area of linguistic meaning, the compositional semantic system could be regarded as capturing what is meaningful, with pragmatic theory capturing tendencies in information structuring and communicative intent; for corpus studies exploring just such a relationship between semantics and pragmatics, see Beaver et al. 2006; Higgins & Sadock 2003; AnderBois et al. 2012.

As recently as 10 years ago, Web search results could also be used to estimate and compare the frequencies of specific words and phrases, but such statistics have become less reliable over the years as a result of a variety of technological and business decisions (Liberman, 2005; Kilgarriff, 2007). To some extent, these needs

can be met with large data distributions like the Google Books project (Michel et al., 2011), but, for the most part, Web searches are reliable only for showing that specific things exist. Robust evidence for statistical tendencies is likely to come only from investigations of stable corpora using tools that allow the analyst to take precise measurements (see sec. 5 for additional comments on methods).

4 Theoretical perspectives on annotations

This volume contains chapters covering best practices in designing annotation schemes, conducting annotation projects, and working with specific corpora. This section is intended as a theorist's companion to those papers. We move from naturalistic annotations like those one might find on the Web to highly focused annotation projects designed to address specific theoretical questions. A recurring theme of our discussion is that annotations are not unimpeachable, but rather the fallible but useful result of interactions among people, machines, and theoretical assumptions.

4.1 Unstructured to highly structured

The most unstructured corpora we consider here are those that are simply collections of raw text, perhaps with document-level divisions given by the structure of the data itself. Such corpora might seem unhelpful for close linguistic analysis, but in fact, once such text is tokenized into (approximations of) linguistically meaningfully units, it can be used to achieve linguistic insights and develop powerful language technologies (Halevy et al., 2009; Norvig, 2009; Turney & Pantel, 2010).

As annotations and other kinds of metadata are added to corpora, they become more richly structured. Because of real-world constraints on time and resources, the more annotations a corpus has, the smaller it is likely to be, but the annotations might enable one to ask more specific and linguistically relevant questions. The most highly structured corpora tend to be those that represent specific interactions like game-play, where the transcript can encode not only what the participants said to each other, but also what they were doing when they said it, what the state of the context was like, and so forth (Thompson et al., 1993; Allen et al., 1996; Stoia et al., 2008; Blaylock & Allen, 2005; Potts, 2012b).

4.2 Naturalistic annotations

If one looks from the right perspective, one finds that the world is full of naturally occurring metadata that can serve as annotations. Such naturalistic annotations tend to be messier than ones created by a trained annotation team, but their super-

abundance can make up for this deficiency. They also have the advantage of being created organically, not as part of a job or artificial task, but rather as part of social, intellectual, and expressive acts that people undertook for their own personal reasons. This can give them a veracity that is often lacking in controlled annotation projects and crowdsourced annotation projects, and it means that they can be studied scientifically in their own right (e.g., Muchnik et al. 2013).

Some annotations are latent in the structure of existing text collections. For example, if one wants to study the language of media bias in the U.S., one might create a corpus of Web data and use the URLs as proxies for political orientation, categorizing `FoxNews.com` as 'right' and `HuffingtonPost.com` as 'left'. Here, the annotations are effectively just (clusters of) addresses. In a similar vein, Thomas et al. (2006) and Monroe et al. (2009) use political speech data, taking the party affiliation of the speaker to be a label for the political orientation of the text. In cases like this, measurement error can be high when compared with what is achievable by hand-labeling, but the vast quantities of available data might make up for this if the theory behind the naturalistic annotations is sound.

At a lower-level, formatting mark-up often encodes valuable clues about linguistic structure. For example, Spitkovsky et al. (2010) and Erlewine (2011) use the boundaries of HTML hyperlink tags as indicators of syntactic constituency, showing that this can help statistical parsing and yield new insights into syntactic structure. (This is another example of complementary insights from NLP and theoretical linguistics; see sec. 2.8.) These cases are of particular interest because they show how features of the text that are not narrowly linguistic can convey information about language structure and content as a by-product of other processes.

The Web also abounds with more explicit metadata intended for business and social networking purposes: 'like' buttons conveying reader reactions, emoticons and hashtags conveying topical and emotional information, 'friend' and 'follower' networks revealing social links, and so forth. The field of sentiment analysis is more or less founded on the notion that star ratings on product and service reviews provide a high-level summary of the attitudes expressed in the associated review text (Pang & Lee, 2005, 2008). At this point, such ratings have been used to train numerous successful sentiment models, for academic and industry purposes, and aspects of the social processes surrounding star ratings have also been studied (Wu & Huberman, 2008). These annotations have to date been less utilized within theoretical linguistics, but see Constant et al. 2009, Potts & Schwarz 2010, and Potts 2011 for attempts to find a role for them in pragmatics.

4.3 Gold-standard annotations

Gold-standard annotations are those that were produced by trained annotators using their linguistic intuitions and a set of guidelines (an annotation manual) to encode implicit structure in a corpus that is not inherently structured along the relevant dimensions. Here, linguists are likely to want to study the annotation manual carefully

to see what concepts it presupposes. In addition, linguists should ask how the final annotations were arrived at. Do they represent an averaging of a number of annotators' judgments? Did the annotators discuss differences and come to a final decision as a committee? And so forth. Most annotation projects report measures of intra-annotator and inter-annotator agreement (Artstein, Takenobu, this volume). Ideally, these are broken down by annotator and category.

It is also important to ascertain whether the annotations themselves match with one's theoretical conception of the issues. On the one hand, one wants consistent, uncontroversial annotations. On the other hand, the pressure to show high agreement could lead to an annotation manual that compromises on crucial theoretical questions or an annotation scheme that masks underlying conceptual muddiness.

What the above amounts to is that the linguist should treat the annotation project as a natural experiment, and the assumptions that went into the experiment should be explicitly represented in the statements of the hypotheses being tested. As an example of where this turned out to be significant, we can contrast the annotations in the FactBank corpus (Saurí 2008; Saurí & Pustejovsky 2009; Saurí, this volume) with the ones obtained by de Marneffe et al. (2012) via crowdsourcing for a subset of the FactBank data. One of the overarching goals of the FactBank annotation project was to encode narrowly semantic intuitions, seeking to factor out pragmatic enrichment deriving from world knowledge and context. The detailed annotation manual emphasizes that the annotators should stay within these bounds. As a result, the annotations conform closely to semantic assumptions but depart from what was intuitively communicated. de Marneffe et al. quantified this intuition with their crowdsourced annotations, which sought to model what was communicated, not what was semantically encoded. Studying the differences between FactBank and these "PragBank" annotations allowed de Marneffe et al. to identify and predict a range of specific kinds of pragmatic enrichment. Stepping back, we see that the nature of these two annotation projects shaped their respective results in theoretically important ways.

4.4 Automatic annotations

Automatic annotations are those that are added by a computer program — for example, one of the widely available part-of-speech taggers, parsers, or named-entity recognizers. These annotations are not guaranteed to be correct or to match any individual speaker's intuitions. Depending on the task, the nature of the model, and the nature of the data, the annotations might be anywhere from near-perfect to completely wrong. Linguists wishing to work with such data should investigate the inferred annotations and become familiar with the patterns of errors. In some cases, the errors will not matter; in others, they will shape the resulting analyses in problematic ways.

To take one complex example, Acton & Potts (To appear) use corpora derived from an online social network to study the social meaning of demonstrative phrases,

as in sentences like *This Henry Kissinger is really something!* and *Make that call right now!* In order to identify demonstrative phrases, they first parsed their data using the Stanford parser with a statistical model trained on newspaper text (Klein & Manning, 2003). The mismatch between the corpus used for training and the one being annotated resulted in many errors,[4] but most were irrelevant to the task at hand; the authors did not need full parses, but rather only a sharp picture of demonstratives. For these, the results were mixed. While the parsing model was basically perfect at identifying demonstrative phrases headed by *this*, *these*, and *those*, it struggled with phrases headed by *that*, which it often confused with complementizer *that* (*We believe that pigs fly*) and relativizer *that* (*the guy that we met*). However, this was not fatal for Acton & Potts's goals: they aimed to study the association between demonstratives and naturalistic annotations in their data, and the non-demonstrative errors for *that* seemed to be fairly evenly distributed across the annotation categories, meaning that their hypothesized demonstrative effect shined through the imperfections in the annotated data, albeit in a weaker form than expected.

4.5 Custom annotation projects

Linguists are apt to ask specialized and focused questions, so custom annotations are often required. As we mentioned above, the mindset of the linguist when working with annotated data should probably resemble the mindset of the psychologist probing experimental results; the nature of the experimental setting (in this case, the annotation project) is every bit as important as the resulting data, and one always wants to study them both together.

In many cases, it is effective for the researchers themselves to annotate their data, especially if the annotations require specialized knowledge. For example, Hacquard & Wellwood (2012) study (among other things) the distribution of epistemic readings of the modal auxiliary *must* in a variety of syntactically embedded contexts. Reliably identifying epistemic readings requires extensive experience with the relevant kinds of data, so the authors made the judgments. The results provide a quantitative picture of the distribution of epistemic modals, and they also exposed the researchers to numerous valuable examples. As is typical for corpus studies, this work confirms a number of hypotheses based on introspection but complicates others.

In our experience (e.g., Harris & Potts 2009; de Marneffe et al. 2008), annotating data oneself offers few savings in terms of time and effort over conducting a full-fledged annotation project involving an annotation team. It does not, for example, obviate the need to have an annotation manual, well-designed annotation interfaces, and tools for studying the resulting annotations to identify errors. Without these things in place, even a lone expert annotator is likely to produce inconsistent, unreliable annotations. This is just to say that it is still important to follow best practices for annotation projects, as covered in other chapters in this volume. In addition, the

[4] For recent attempts to build tagging and parsing models that are better-suited to informal Web data, see Ritter et al. 2011; Owoputi et al. 2013; de Marneffe et al. 2013.

linguist annotating the data himself should be careful to avoid theoretical biases, perhaps restricting self-annotation to scenarios where he has no vested stake in any particular result, but rather is seeking to use the corpus to help discover patterns, say, to inform a psycholinguistic experiment.

At present, crowdsourcing platforms make it relatively easy to get custom annotations for specialized tasks. As with regular human-subjects experiments, crowdsourcing is limited by what people can do with little or no training, but scientists throughout the cognitive and computational sciences have shown that incredible work can be done despite this limitation (Snow et al., 2008; Callison-Burch, 2009; Heer & Bostock, 2010; Hsueh et al., 2009; Munro et al., 2010; Sprouse, 2010). Rather than trying to survey this large literature (see Poesio, this volume), we want to highlight two novel uses that theoreticians might make of crowdsourcing.

First, crowdsourcing paves the way to getting a large number of annotations for each example and studying the results the way one would study response data from a questionnaire-based experiment. Although the norm in crowdsourcing is to collect just 3–5 responses per example and use the majority choice as the true annotation, it is often possible to collect upwards of 20 responses per example, meaning that one can study the variance in the response distributions and use statistical tests to assess the reliability of the resulting annotation. With such corpora, the analyst can choose a majority annotation (where there is one), perhaps associated with a measure of uncertainty, or else just treat each example as labeled with its full response distribution (de Marneffe et al., 2012).

Second, crowdsourced data can be explicitly or implicitly interactional in a way that traditional corpus annotations are not. For example, Potts (2012b) reports on the publicly available Cards corpus, which consists mainly of transcripts of Amazon Mechanical Turk workers playing an interactive chat game with each other in real time. Alternatively, the interactional component could be implicit, a part of the instructions given to the annotators, guided by a theory of interaction and communication. Clarke et al. (2013) created and released a corpus to investigate how visual salience impacts the production of referring expressions. The workers were asked to describe a target so that someone else could find it in a complex visual scene. Before acting as producers, they were placed in the role of interpreter, by completing a training phase designed to increase their awareness of how ambiguities are perceived. The resulting multi-modal corpus opens the door to further study of how visual features interact with semantic and pragmatic features. For example, Duan et al. (2013) use the corpus to study how visual salience influences the definiteness of referring expressions.

We have found that crowdsourcing is a powerful technique for getting custom annotations, and the annotation phase is typically much faster than for traditional annotation projects. However, these gains should be weighed against the time and effort it takes to set up a successful crowdsourcing experiment and interpret the results. Regarding set-up, crowdsourcing requires all of the care and attention of a psycholinguistic experiment, and taking shortcuts will lead to poor results and unhappy workers. Regarding interpretation, crowdsourced annotations are likely to have higher variance than traditional annotation projects, even taking into account

the larger numbers of people involved. Crowdsourcing is often touted as a fast route to annotations, but the reality is more nuanced, with expert annotations proving easier in many circumstances.

5 Methods and modes of inquiry

This section outlines the basic steps involved in conducting corpus work, from data wrangling to hypothesis formation and testing. We can't offer lock-step advice because, like all scientific inquiry, the specific steps will be particular to the research questions and will be deeply entwined with the specialized knowledge of the researchers themselves. We mainly aim to highlight the ways in which the methods and modes of inquiry are part of the scientific project itself.

5.1 Programming basics

The corpus linguists of the late 19th and early 20th century painstakingly tabulated frequencies by hand. Working in the early 1960s, Francis & Kučera (1964) were only slightly more computationally fortunate, typing the now-famous Brown corpus onto punch cards (Kučera & Francis, 1967; Francis & Kučera, 1979). By contrast, the linguists of the early 21st century have it easy. Modern programming languages have removed all the major barriers to doing advanced computational analysis; in our experience as teachers, it takes just a few weeks of guided coding and practice for students to go from having no programming experience to doing sophisticated analysis on large corpora. While one can accomplish a lot with Web searches and basic spreadsheet programs, learning a programming language is easy, empowering, and increasingly a basic part of scientific literacy.

Our sense is that, at the time of this writing, the dominant programming languages for corpus linguistics are Java,[5] Python,[6] and R.[7] These languages are freely available, easy to use, and powerful. Their dominance within linguistics also owes in part to the excellent textbooks and computational libraries written for them, including the Stanford NLP tools (Klein & Manning, 2003; Toutanova et al., 2003; Finkel et al., 2005; Lee et al., 2011; Recasens et al., 2013), Python NLTK (Bird et al., 2009), and the languageR package (Baayen, 2008). In our view, Java and Python are currently the better choices for doing heavy-duty text processing, R is currently the best choice for doing statistical analysis and visualization, and Python and R are better for writing small programs ('scripting') and deploying them quickly. However, the differences are rapidly disappearing (Gries, 2009; Odersky et al., 2010; McKinney,

[5] `http://java.com/`

[6] `http://www.python.org`

[7] `http://www.r-project.org`

2012), and software has been written to make language-processing functions available in each of these languages available in the others, so we think aspiring and experienced corpus linguists alike will be well-served by any of them.[8]

5.2 Getting to know your corpus

The title of this section is taken from Kilgarriff (2012), who encourages the linguist working with a new corpus to undertake lots of informal fact-finding missions as part of the cycle of developing and testing hypotheses.

Ideally, one would read the entire corpus through, on the look-out for idiosyncrasies. However, modern corpora tend to be so large as to make a deep read impractical. In such cases, we still advise reading samples, both randomly and strategically. For annotated data, this sampling can be done effectively in conjunction with studying the annotation manual, as a way of getting inside the minds of the annotators themselves. However, Kilgarriff and also Fillmore (1992) emphasize that close reading has weaknesses as well as strengths. It is likely to provide the reader with a deep understanding of the content of the texts, and perhaps glimpses into the underlying contexts and social forces, but it is unlikely to reveal unexpected distributions in linguistic units, hard-to-see encoding inconsistencies, systematic annotation errors (sec. 4.4), and other phenomena that require wide-scale statistical analysis or the finicky inflexibility that only a computer program can guarantee.

Thus, reading in the usual (human) sense is always fruitfully paired with wide-scale computational analysis: creating word lists and sorting them by frequency, visualizing the distribution of word frequencies (Baayen 2001:§1), studying the distributions of any metadata contained in the corpus (usernames, dates, locations, ratings, etc.), relating the metadata distributions to each other and to the language data, and so forth. This process inevitably turns up oddities of the underlying corpus, reveals shortcomings in one's code for processing the corpus, and, more positively, helps in aligning one's hypotheses with the corpus. Data analysis experts in many fields tend to value visualization over statistical analysis at this stage, since it can often tell a more complex story and is less likely to hide assumptions that might be problematic; for discussion and advice on best practices, see Cleveland 1985; Tufte 2001; Baayen 2008; Chen et al. 2008; Gries 2009.

5.3 Feature extraction

In computational linguistics and NLP, *feature extraction* is the task of identifying, isolating, and clustering units from a data collection that are meaningful for the

[8] For phonetic analysis, all these languages still lag behind Praat (Boersma & Weenink, 2013).

analysis. This step always involves a mix of theoretical assumptions and heuristic approximations, and is thus a central piece of any corpus analysis.

To see that things can get very complex very quickly, consider a hypothetical corpus study aimed at studying the relative frequency of different weather verbs like *snow*, *sleet*, and *hail*. The intuitive feature extraction task is just to identify these verbs for the purpose of counting them by type. The actual feature extraction function will involve numerous non-trivial choices. Which verbs should be included in the input list? Should morphological tense variants (e.g., *snow*, *snows*, *snowed*) be collapsed together? What about aspectual forms (e.g., *snowing*)? Are metaphorical uses (*The problem snowed me*) frequent enough that they need to be addressed separately? How will verbal uses be distinguished from others — does the corpus have gold-standard part-of-speech tags, or will these need to be automatically assigned? In the case of automatic assignment, are there weather-verb-related biases we should know about (e.g., an overwhelming bias against analyzing *blizzard* as a verb even where that is the correct choice).

We could go on, and we have hardly even touched on the issue of how these choices interact with the nature of the corpus itself (weather reports in Finland, Germany, Egypt?). As the research question gets more complex, the number of choices tends to grow quickly. This can be worrying or freeing, depending on the perspective one takes. On the one hand, one might worry about the implications for scientific validity. *Researcher degree of freedom* is a primary concern for scientific research in general: if the researcher is allowed to modify his hypotheses and methods until the analysis 'works', then basic statistical principles lead us to expect a lot of spurious conclusions (Simmons et al., 2013). On the other hand, because of the nature of corpus research, it is typically be possible for the researcher to release every aspect of his analysis to the public: not only the data, but also the functions used for feature extraction and analysis. Whereas only some of the details can be included in the official research report, the code can expose everything, allowing others to directly reproduce reported results and explore alternatives. This puts pressure on the scientist, but in a way that we can all regard as intellectually healthy.

5.4 Forming specific hypotheses

We saw in secs. 3–4 that dealing with corpus data and annotations can be a delicate matter. We now seek to connect those observations explicitly with hypothesis formulation and testing. In our experience, the process typically involves moving from an intuitive hypothesis about language to a technical hypothesis about particular corpora and annotations, in much the same way that psycholinguists move from theory to experimental design. Framing one's investigations in these terms might seem cumbersome, but it can be productive: it facilitates testing the same intuitive hypothesis with multiple diverse corpora, and it creates opportunities to scrutinize not only the intuitive hypothesis but also its relationship to the technical one.

As an example, consider the intuitive hypothesis that prepositions in English cannot have finite clausal complements. This hypothesis entails, for example, that *we boasted about the fact that we won* is grammatical, whereas *we boasted about that we won* is ungrammatical. Suppose we are working with the Penn Treebank 3 (Marcus et al., 1999), which contains gold-standard parse trees for the Brown corpus (Kučera & Francis, 1967), newspaper data, and the Switchboard conversational corpus (Godfrey & Holliman, 1997). Then our technical hypothesis will be given in terms of a set of subtrees (bracketed strings) that we identify with regular expressions (Friedl, 2006; Levy & Andrew, 2006). Call this set of trees *S*.

It is tempting to say that the hypothesis is simply that no member of *S* occurs in the treebank. However, as we saw in sec. 3, this probably will not suffice. Suppose the corpus does contain a member of *S*. What are the chances that this observation is due to the interaction of irrelevant factors like disfluencies in speech, typographic errors in print, or annotation mistakes? Conversely, suppose the corpus does not contain a member of *S*. Setting aside the possibility of simple experimental error, how confident can we be that this is truly indicative of a linguistic constraint?

These realities suggest that the technical hypothesis is best stated in statistical terms, even if the intuitive hypothesis is categorical. For example, the hypothesis might say that the ratio of nominal prepositional objects to clausal prepositional objects is vanishingly small, even taking into consideration the frequencies of the relevant constituents. To account for the possibility that the data actually contain no clausal prepositional objects, we might adopt a model-based approach to calculating the relevant values, to avoid tailoring our measurement too closely to the treebank itself (Pereira, 2000; Domingos, 2012), which is, after all, just a source of evidence, not our ultimate object of study.

The probabilistic nature of corpus evidence encourages a further encoding of one's hypotheses using the language of statistical hypothesis testing. This can be useful analytically, and it helps in getting results accepted by the scientific community. However, in addition to the usual concerns about using statistical tests in this way (Gelman & Stern, 2006), corpus data present at least two special challenges. First, in large, naturalistic corpora, there are typically so many unmeasured interacting factors that the null hypothesis being tested tends to be trivially false and is, at any rate, not of real interest (Kilgarriff 2005; cf. Gries 2005). Second, word distributions are unusual in nature (Zipf, 1949; Baayen, 2001), so most parametric statistical tests implicitly depend on distributional assumptions that are false of the raw corpus data.

This is not to say that statistical hypothesis testing is always uninformative for corpus data. It can certainly help with decision making, especially where one can show large effect sizes and stable results across samples from the full dataset. Hypothesis testing can also be supplemented by evaluations on new data, using the train–development–test methodology that dominates NLP. Such evaluations provide information about the practical significance of the hypotheses and help to avoid conclusions that are tailored to the particular corpus at hand. Above all else, though, we advise having specific, well-articulated theoretical motivations for one's hypotheses going in. In the context of theoretical work, rich and specific connections with the

literature are likely to carry the most weight within the community, and they are the best way to ensure that the necessary exploratory data analysis is productive rather than insidious.

6 Conclusion

Corpus linguists and theoretical linguists once took themselves to be locked in a bitter debate about the foundations of linguistic theory and the proper conduct of linguistic investigations. We won't repeat the epithets here. Both sides seem to have emerged triumphant. Fillmore (1992) self-identifies as a "computer-assisted armchair linguist". We also know experiment-assisted corpus linguists, computer-assisted psycholinguists, experiment-assisted armchair linguists, armchair-assisted psycholinguists, and armchair-assisted corpus linguists. In the end, we expect all of these titles to reduce to 'linguist'. Our central argument is that corpus, introspective, and psycholinguistic methods all complement each other; far from being in tension methodologically or philosophically, they can be brought together to strengthen linguistic theory and increase its scope and scientific relevance.

References

Acton, Eric K. & Christopher Potts. To appear. That straight talk: Sarah Palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics* .

Allen, James F., Bradford W. Miller, Eric K. Ringger & Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *Proceedings of the 34th annual meeting of the Association for Computational Linguistics*, 62–70. Santa Cruz, CA: ACL.

AnderBois, Scott, Adrian Brasoveanu & Robert Henderson. 2012. The pragmatics of quantifier scope: A corpus study. In Ana Aguilar-Guevara, Anna Chernilovskaya & Rick Nouwen (eds.), *Proceedings of Sinn und Bedeutung 16*, vol. 1 MIT Working Papers in Linguistics, 15–28. Cambridge, MA: MIT Linguistics.

Andor, József. 2004. The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1(1). 93–111.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics*. Cambridge University Press.

Barton, Stephen B. & Anthony J. Sanford. 1993. A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory and Cognition* 21(4). 477–487.

Beaver, David I. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy* 27(1). 3–56.

Beaver, David I. 2007. Corpus pragmatics: Something old, something new. Paper presented at the annual meeting of the Texas Linguistic Society.

Beaver, David I., Itamar Francez & Dmitry Levinson. 2006. Bad subject! (Non)-canonicality and NP distribution in existentials. In Effi Georgala & Jonathan Howell (eds.), *Proceedings of Semantics and Linguistic Theory 15*, 19–43. Ithaca, NY: CLC Publications.

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media.

Blaylock, Nate & James F. Allen. 2005. Generating artificial corpora for plan recognition. In Liliana Ardissono, Paul Brna & Antonija Mitrovic (eds.), *User modeling 2005* Lecture Notes in Artificial Intelligence, 179–188. Berlin: Springer.

Bock, Kathryn, Sally Butterfield, Anne Cutler, J. Cooper Cutting, Kathleen M. Eberhard & Karin R. Humphreys. 2006. Number agreement in British and American English: Disagreeing to agree collectively. *Language* 82(1). 64–113.

Bod, Rens, Jennifer Hay & Stefanie Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Boersma, Paul & David Weenink. 2013. Praat: Doing phonetics by computer. Computer program; version 5.3.60. `http://www.praat.org/`.

Bresnan, Joan & Tatiana Nikitina. 2010. The gradience of the dative alternation. In Linda Uyechi & Lian Hee Wee (eds.), *Reality exploration and discovery: Pattern interaction in language and life*, 161–184. Stanford, CA: CSLI.

Burge, Tyler. 1979. Individualism and the mental. In Peter French, Theodore Uehling & Howard Wettstein (eds.), *Midwest studies in philosophy*, vol. IV: Studies in Metaphysics, 73–121. Minneapolis: University of Minnesota Press.

Callison-Burch, Chris. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, 286–295. Singapore: ACL.

Chen, Chun-houh, Wolfgang Karl Härdle & Antony Unwin (eds.). 2008. *Handbook of data visualization*. Berlin: Springer.

Chomsky, Noam. 1957a. A review of B. F. Skinner's *Verbal Behavior*. *Language* 35(1). 26–58.

Chomsky, Noam. 1957b. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam. 1986. *Knowledge of language*. New York: Praeger.

Clark, Herbert H. 1997. Dogmas of understanding. *Discourse Processes* 23(3). 567–59.

Clarke, Alasdair D. F., Micha Elsner & Hannah Rohde. 2013. Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology (Perception Science)* 4(1). 1–10.

Cleveland, William S. 1985. *The elements of graphing data*. Summit, NJ: Hobart Press.

Constant, Noah, Christopher Davis, Christopher Potts & Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung* 33(1–2). 5–21.

Cover, Thomas M. & Joy A. Thomas. 1991. *Elements of information theory*. New York: Wiley.

Culbertson, Jennifer & Steven Gross. 2009. Are linguists better subjects? *The British Journal for the Philosophy of Science* 60(4). 721–736.

Degen, Judith. 2013. A corpus-based study of *Some* (but not *All*) implicatures. Ms., University of Rochester.

Devitt, Michael. 2006. Intuitions in linguistics. *The British Journal for the Philosophy of Science* 57(3). 481–513.

Dewey, Godfrey. 1923. *Relative frequency of English speech sounds*. Harvard University Press.

Domingos, Pedro. 2012. A few useful things to know about machine learning. *Communications of ACM* 55(10). 78–87.

Duan, Manjuan, Micha Elsner & Marie-Catherine de Marneffe. 2013. Visual and linguistic predictors for the definiteness of referring expressions. In *Proceedings of the 17th workshop on the semantics and pragmatics of dialogue*, 25–34.

Erlewine, Michael Yoshitaka. 2011. The constituency of hyperlinks in a hypertext corpus. Ms., MIT.

Faye, Jan. 2008. Copenhagen interpretation of quantum mechanics. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, CSLI fall 2008 edition edn. `http://plato.stanford.edu/archives/fall2008/entries/qm-copenhagen/`.

Fillmore, Charles J. 1992. "Corpus linguistics" or "computer-aided armchair linguistics". In Svartvik (1992) 35–66.

Finkel, Jenny Rose, Trond Grenager & Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 363–370. Ann Arbor, MI: ACL.

Francis, W. Nelson & Kučera. 1979. Manual of information to accompany a 'standard sample of present-day edited American English, for use with digital computers'. Tech. rep. Brown University Providence, RI.

Francis, W. Nelson & Henry Kučera. 1964. A standard sample of present-day English for use with digital computers. Report to the U. S. Office of Education on Cooperative Research Project E-007 Brown University Providence, RI.

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the Cognitive Science Society*, 939–944. Washington, D.C.

Frazier, Lyn. 2012. Co-reference and adult language comprehension. *Revista Linguistica* 8(2). 1–11.

Friedl, Jeffrey E. F. 2006. *Mastering regular expressions*. Sebastopol, CA: O'Reilly Media 3rd edn.

Gelman, Andrew. 2011. Review essay: Causality and statistical learning. *American Journal of Sociology* 117(3). 955–966.

Gelman, Andrew & Hal S. Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician* 60(4). 328–331.

Glass, Lelia. 2013. What does it mean for an implicit object to be recoverable? In *Proceedings of the Penn linguistics colloquium*, Philadelphia, PA: Penn Linguistics Club.

Godfrey, John J. & Ed Holliman. 1997. Switchboard-1 release 2. Linguistic Data Consortium, Catalog #LDC97S62.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2). 153–198.

Goldwater, Sharon, Thomas L. Griffiths & Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics*, 673–680. Sydney, Australia: ACL.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Stockholm workshop on variation within Optimality Theory*, 111–120. Stockholm: Stockholm University.

Goodman, Noah D. & Daniel Lassiter. To appear. Probabilistic semantics and pragmatics. In Shalom Lappin & Chris Fox (eds.), *The handbook of contemporary semantic theory*, Oxford: Wiley-Blackwell 2nd edn.

Gordon, Peter C., Barbara J. Grosz & Laura A. Gilliom. 1993. Pronouns, names and the centering of attention in discourse. *Cognitive Science* 17(3). 311–348.

Gordon, Peter C. & Randall Hendrick. 1997. Intuitive knowledge of linguistic coreference. *Cognition* 3(3). 325–370.

Gries, Stefan Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2). 277–294.

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge.

Grimm, Scott & Louise McNally. 2013. No ordered arguments needed for nouns. In Maria Aloni, Michael Franke & Floris Roelofsen (eds.), *Proceedings of the 19th Amsterdam colloquium*, 123–130. Amsterdam: ILLC.

Hacquard, Valentine & Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5(4). 1–29.

Halevy, Alon, Peter Norvig & Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24(2). 8–12.

Harris, Jesse A. & Christopher Potts. 2009. Perspective-shifting with appositives and expressives. *Linguistics and Philosophy* 32(6). 523–552.

Harris, Randy Allen. 1993. *The linguistic wars*. Oxford: Oxford University Press.

Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146–162.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.

Heer, Jeffrey & Michael Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *ACM human factors in computing systems*, 203–212.

Higgins, Derrick & Jerrold M. Sadock. 2003. A machine learning approach to modeling scope preferences. *Computational Linguistics* 29(1). 73–96.

Hockett, Charles F. 1948. A note on 'structure' [review of de Goeje by W. D. Preston]. *International Journal of American Linguistics* 14(4). 269–171.

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10(2). 210–234.

Hoeksema, Jack. 1997. Corpus study of negative polarity items. University of Groningen. `http://www.let.rug.nl/hoeksema/docs/barcelona.html`.

Hoeksema, Jack. 2008. There is no number effect in the licensing of negative polarity items: A reply to Guerzoni and Sharvit. *Linguistics and Philosophy* 31(4). 397–407.

Hofmeister, Philip & Ivan A. Sag. 2010. Cognitive constraints and island effects. *Language* 22(6). 366–415.

Hofmeister, Philip, Laura Staum Casasanto & Ivan A. Sag. 2012a. How do individual cognitive differences relate to acceptability judgments? A reply to Sprouse, Wagers, and Phillips. *Language* 88(2). 390–400.

Hofmeister, Philip, Laura Staum Casasanto & Ivan A. Sag. 2012b. Misapplying working-memory tests: A reductio ad absurdum. *Language* 88(2). 408–409.

Horn, Laurence R. 1991. Duplex negatio affirmat. . . : The economy of double negation. In Lise M. Dobrin, Lynn Nichols & Rosa M. Rodriguez (eds.), *Papers from the 27th regional meeting of the Chicago Linguistic Society*, vol. 2: The Parasession on Negation, 80–106. Chicago: Chicago Linguistic Society.

Hsueh, Pei-Yun, Prem Melville & Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, 27–35. Boulder, CO: ACL.

Jackendoff, Ray S. 1992. *Languages of the mind*. Cambridge, MA: MIT Press.

Jurafsky, Dan. 1996. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20(2). 137–194.

Jurafsky, Daniel & James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Englewood Cliffs, NJ: Prentice-Hall 2nd edn.

Katz, Jerrold J. 1981. *Language and other abstract objects*. Totowa, NJ: Rowman and Littlefield.

Katz, Jerrold J. & Paul M. Postal. 1991. Realism vs. conceptualism in linguistics. *Linguistics and Philosophy* 14(5). 515–554.

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics* 33(1). 147–151.

Kilgarriff, Adam. 2012. Getting to know your corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, speech and dialogue: 15th international conference*, vol. 7499 Lecture Notes in Artificial Intelligence, 3–15. Berlin: Springer.

Kilgarriff, Adam & Edward Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29(3). 333–347.

Klein, Dan & Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, vol. 1, 423–430. Sapporo, Japan: ACL.

Kučera, Henry & W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kwiatkowski, Tom, Luke S. Zettlemoyer, Sharon Goldwater & Mark Steedman. 2011. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the conference on empirical methods in natural language processing*, 1512–1523. Edinburgh: ACL.

Lassiter, Daniel. 2008. Semantic externalism, language variation, and sociolinguistic accommodation. *Mind and Language* 23(5). 607–633.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu & Daniel Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the 15th conference on computational natural language learning: Shared task*, 28–34. Portland, OR: ACL.

Leech, Geoffrey N. 1992. Corpora and theories of linguistic performance. In Svartvik (1992) 105–122.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: Chicago University Press.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177.

Levy, Roger & Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation*, 2231–2234.

Levy, Roger & T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John Platt & Thomas Hoffman (eds.), *Advances in neural information processing systems 19*, 849–856. Cambridge, MA: MIT Press.

Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press. Reprinted 2002 by Blackwell.

Liang, Percy, Michael I. Jordan & Dan Klein. 2013. Learning dependency-based compositional semantics. *Computational Linguistics* 39(2). 389–446.

Liberman, Mark. 2005. Questioning reality. Language Log, January 24. `http://itre.cis.upenn.edu/~myl/languagelog/archives/001837.html`.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates 3rd edn.

Manning, Christopher D. 2003. Probabilistic syntax. In Bod et al. (2003) 289–341.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, Mitchell P., Beatrice Santorini, Mary A. Marcinkiewicz & Ann Taylor. 1999. The Penn treebank 3. Linguistic Data Consortium, Catalog #LDC99T42.

de Marneffe, Marie-Catherine, Miriam Connor, Natalia Silveira, Samuel R. Bowman, Timothy Dozat & Christopher D. Manning. 2013. More constructions, more genres: Extending Stanford Dependencies. In Eva Hajičová, Kim Gerdes & Leo Wanner (eds.), *Proceedings of the second international conference on dependency linguistics*, 187–196. Prague: ACL.

de Marneffe, Marie-Catherine, Christopher D. Manning & Christopher Potts. 2010. "Was it good? It was provocative." Learning the meaning of scalar adjectives. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 167–176. Uppsala, Sweden: ACL.

de Marneffe, Marie-Catherine, Christopher D. Manning & Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics* 38(2). 301–333.

de Marneffe, Marie-Catherine, Anna N. Rafferty & Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th annual meeting of the Association for Computational Linguistics*, 1039–1047. Columbus, OH: ACL.

McEnery, Tony & Andrew Wilson. 2001. *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

McKinney, Wes. 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. Sebastopol, CA: O'Reilly Media.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.

Monroe, Burt L, Michael P. Colaresi & Kevin M. Quinn. 2009. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4). 372–403.

Muchnik, Lev, Sinan Aral & Sean J. Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341(6146). 647–651.

Munro, Rob. 2012. *Processing short message communications in low-resource languages*. Stanford, CA: Stanford University dissertation.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, 122–130. Los Angeles: ACL.

Norvig, Peter. 2009. Natural language corpus data. In Toby Segaran & Jeff Hammerbacher (eds.), *Beautiful data*, 219–242. O'Reilly Media.

Norvig, Peter. 2011. On Chomsky and the two cultures of statistical learning. Google, Inc. `http://norvig.com/chomsky.html`.

Odersky, Martin, Lex Spoon & Bill Venners. 2010. *Programming in Scala*. Walnut Creek, CA: Artima 2nd edn.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider & Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of*

the North American chapter of the Association for Computational Linguistics: Human language technologies, 380–390. Atlanta, GA: ACL.

Pang, Bo & Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 115–124. Ann Arbor, MI: ACL.

Pang, Bo & Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1). 1–135.

Pereira, Fernando C. N. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society* 358(1769). 1239–1253.

Potts, Christopher. 2011. On the negativity of negation. In Nan Li & David Lutz (eds.), *Proceedings of Semantics and Linguistic Theory 20*, 636–659. Ithaca, NY: CLC Publications.

Potts, Christopher. 2012a. Conventional implicature and expressive content. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 3, 2516–2536. Berlin: Mouton de Gruyter.

Potts, Christopher. 2012b. Goal-driven answers in the Cards dialogue corpus. In Nathan Arnett & Ryan Bennett (eds.), *Proceedings of the 30th West Coast Conference on Formal Linguistics*, 1–20. Somerville, MA: Cascadilla Press.

Potts, Christopher & Florian Schwarz. 2010. Affective 'this'. *Linguistic Issues in Language Technology* 3(5). 1–30.

Putnam, Hilary. 1975. *Mind, language, and reality: Philosophical papers*, vol. 2. Cambridge: Cambridge University Press.

Rajaraman, Anand & Jeffrey D. Ullman. 2011. *Mining of massive datasets*. Cambridge: Cambridge University Press.

Recasens, Marta, Marie-Catherine de Marneffe & Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Human language technologies: The 2013 annual conference of the North American chapter of the Association for Computational Linguistics*, 627–633. Atlanta, Georgia: ACL.

Ring, Nicholas & Alexandra L. Uitdenbogerd. 2009. Finding 'Lucy in disguise': The misheard lyric matching problem. In Gary Geunbae Lee, Dawei Song, Chin-Yew Lin, Akiko Aizawa, Kazuko Kuriyama, Masaharu Yoshioka & Tetsuya Sakai (eds.), *Information retrieval technology: 5th Asia information retrieval symposium* (Lecture Notes in Computer Science 5839), 157–167. Berlin: Springer.

Ritter, Alan, Sam Clark, Mausam & Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, 1524–1534. Edinburgh: ACL.

Roark, Brian & Richard Sproat. 2007. *Computational approaches to morphology and syntax*. Cambridge, MA: Oxford University Press.

Sag, Ivan A. & Thomas Wasow. 2011. Performance-compatible competence grammar. In Robert Borsley & Kersti Börjar (eds.), *Non-transformational syntax: Formal and explicit models of grammar*, 359–377. Oxford: Wiley-Blackwell.

Saurí, Roser. 2008. *A factuality profiler for eventualities in text*: Computer Science Department, Brandeis University dissertation.

Saurí, Roser & James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation* 43(3). 227–268.

Scholz, Barbara C., Francis Jeffry Pelletier & Geoffrey K. Pullum. 2011. Philosophy of linguistics. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Stanford, CA: CSLI winter 2011 edn. `http://plato.stanford.edu/archives/win2011/entries/linguistics/`.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Schütze, Carson T. 2009. Web searches should supplement judgements, not supplant them. *Zeitschrift für Sprachwissenschaft* 28(1). 151–156.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423, 623–656.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2013. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11). 1359–1366.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Y. Ng. 2008. Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263. Honolulu, Hawaii: ACL.

Snyder, William. 2000. An experimental investigation of syntactic satiation effects. *Linguistic Inquiry* 31(3). 575–582.

Spencer, N. J. 1973. Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of Psycholinguistic Research* 2(2). 83–98.

Spitkovsky, Valentin I., Daniel Jurafsky & Hiyan Alshawi. 2010. Profiting from mark-up: Hyper-text annotations for guided parsing. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 1278–1287. Uppsala, Sweden: ACL.

Sproat, Richard & Chilin Shih. 1991. The cross-linguistic distribution of adjective ordering restrictions. In Carol Georgopoulos & Roberta Ishihara (eds.), *Interdisciplinary approaches to language: Essays in honor of S.-Y. Kuroda*, 565–593. Berlin: Springer.

Sprouse, Jon. 2010. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167.

Sprouse, Jon & Norbert Hornstein (eds.). 2013. *Experimental syntax and the islands debate*. Cambridge: Cambridge University Press.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134. 219–248.

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012a. A test of the relation between working memory capacity and syntactic island effects. *Language* 88(1). 82–123.

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012b. Working-memory capacity and island effects: A reminder of the issues and the facts. *Language* 88(2). 401–407.

Stoia, Laura, Darla Magdalene Shockley, Donna K. Byron & Eric Fosler-Lussier. 2008. SCARE: A situated corpus with annotated referring expressions. In *Proceedings of the 6th international conference on language resources and evaluation*, Marrakesh, Morroco: European Language Resources Association.

Svartvik, Jan (ed.). 1992. *Directions in corpus linguistics: Proceedings of Nobel symposium 82*. Berlin: Mouton de Gruyter.

Thomas, Matt, Bo Pang & Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, 327–335. Sydney, Australia: ACL.

Thompson, Henry S., Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands & Cathy Sotillo. 1993. The HCRC map task corpus: Natural dialogue for speech recognition. In *HLT '93: Proceedings of the workshop on human language technology*, 25–30. Princeton, NJ: ACL.

Thuilier, Juliette, Anne Abeille & Benoît Crabbé. 2013. Ordering preferences for postverbal complements in French. In Henry Tyne, Virginie André, Alex Boulton & Christophe Benzitoun (eds.), *Ecological and data-driven perspectives in French language studies*, Cambridge: Cambridge Scholars Publishing.

Toutanova, Kristina, Dan Klein, Christopher D. Manning & Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics*, vol. 1 NAACL '03, 173–180. Edmonton, Canada: ACL.

Tufte, Edward R. 2001. *The visual display of quantitative information*. Cheshire, CT: Graphics Press 2nd edn.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37. 141–188.

Vickers, John. 2013. The problem of induction. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, CSLI spring 2013 edn. `http://plato.stanford.edu/entries/induction-problem/`.

Vitevitch, Michael S. 2002. Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear. *Language and Speech* 45(4). 407–434.

Walker, Marilyn A., Aravind K. Joshi & Ellen F. Prince (eds.). 1997. *Centering in discourse*. Oxford University Press.

Wason, Peter C. & Shuli S. Reich. 1979. A verbal illusion. *Quarterly Journal of Experimental Psychology* 31(4). 591–597.

Wierzbicka, Anna. 1987. *English speech act verbs: A semantic dictionary*. New York: Academic Press.

Winston, Andrew S. & Daniel J. Blais. 1996. What counts as an experiment? A transdisciplinary analysis of textbooks, 1930–1970. *The American Journal of Psychology* 109(4). 599–616.

Wong, Yuk Wah & Raymond Mooney. 2007. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th annual meet-*

ing of the Association for Computational Linguistics, 960–967. Prague, Czech Republic: ACL.

Wu, Fang & Bernardo A. Huberman. 2008. How public opinion forms. In Christos Papadimitriou & Shuzhong Zhang (eds.), *Internet and network economics*, vol. 5385 Lecture Notes in Computer Science, 334–341. Berlin: Springer.

Zettlemoyer, Luke S. 2009. *Learning to map sentences to logical form*. Cambridge, MA: MIT dissertation.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.