A Simple Method for Tagset Comparison

Markus Dickinson and Charles Jochim

Indiana University
md7@indiana.edu, cajochim@indiana.edu

Abstract

Based on the idea that local contexts predict the same basic category across a language, we develop a simple method for comparing tagsets across corpora. The principle differences between tagsets are evidenced by variation in categories in one corpus in the same contexts where another corpus exhibits only a single tag. Such mismatches highlight differences in the definitions of tags which are crucial when porting technology from one annotation scheme to another.

1. Introduction and Motivation

While it is desirable to develop corpus-independent methods in natural language processing for learning from corpus annotation, the distinctions made in corpus annotation can have a great impact on accuracy. For example, which part-of-speech (POS) distinctions are made affects the success of a tagger (cf., e.g., Brants, 1997), especially the amount of non-locality in tagging distinctions across tagsets. Tagging algorithms thus sometimes use tagset-specific features or features optimized for a particular annotation scheme (e.g., Toutanova and Manning, 2000). The difficulty then lies in transferring the technology from one annotation scheme to another. For example, if one annotation scheme does not explicitly annotate case, technology designed to be trained on that scheme may have difficulty in moving to a scheme which does, in that it may not have been designed to emphasize morphological properties. Finding and understanding such differences is thus crucial for improving technology, in addition to providing a window on better linguistic understanding.

To determine the exact effect of the tagset on technology, one must compare annotation schemes. Specifically, one needs to investigate the quality of a (morphosyntactic) tagset, or annotation scheme, where by *quality* we refer to the linguistic information encoded in a tagset, and which can vary in syntactically meaningful ways across tagsets. Two tagsets can differ in quality if one encodes distinctions that the other does not, e.g., a difference in baseform and present tense verbs.

Tagset comparison tends to either emphasize the *internal* quality of a tagset, i.e., whether it can be tagged accurately, or the *external* quality, i.e., whether it makes important linguistic distinctions (see Déjean, 2000, sec. 2 & 7), and such methods generally either require sophisticated machinery or complete manual evaluation. What would be preferable is an automatic or semi-automatic method to compare tagsets which uses only simple properties from the data. Further, we want it to tell us something about what the categories denote across tagsets, both to pinpoint the most problematic cases for taggers and to provide feedback to humans (cf. also Babarczy et al., 2006).

A simple question that provides some guidance in exploring these issues is: for cases where two annotation schemes could have made the same tagging decision, do they in fact make the same distinction? Or do they differ? Note that this question sidesteps the issue of internal and external quality, instead focusing on a more general notion of quality, encompassing both. The key to answering this question lies in defining properties of different corpora such that we can determine when the same morphosyntactic decision could have been made.

A field of research that defines such properties, often relying on simple but accurate corpus properties, is annotation error detection. Techniques for detecting inconsistencies predict areas of consistent annotation. Furthermore, the method we build from (Dickinson and Meurers, 2003) uses the local contextual information shown to be relevant for human category learning (Mintz, 2003). Thus, as described below, we have a simple and cognitively plausible method to use, which gives an indication about the quality of the categories.

The representation of data for error detection, then, allows for a relatively simple way to compare annotations. As discussed in section 2., very short contexts can be found across corpora, and thus inconsistency detection can highlight areas where the context is the "same," thereby predicting similar tags across different corpora. By using only local context, differences in tags indicate that either there is a potentially unclear distinction which humans have difficulty with, or the distinction requires non-local information to disambiguate. The latter is a problem for taggers (internal criterion), and the former a problem in the category definition (external criterion). As outlined in section 3.1., this cross-corpus comparison also provides a more thorough evaluation of corpus-independent error detection methods. In section 3.2., we discuss using the same representation to more robustly pinpoint tagset discrepancies, and in section 4. we evaluate this work.

2. Comparable units of data

To compare tagsets, we need units of data to compare across corpora that allow for the same tagging decision to be made. But what are comparable contexts? To find a suitable definition, we start with the variation n-gram error detection method (Dickinson and Meurers, 2003; Dickinson, 2005), which detects errors by looking for items occurring multiple times in a corpus with varying annotation. These *variation nuclei* are predicted to be unambiguous when a part of the same *variation n-gram*. For example, in the

Wall Street Journal (WSJ) corpus, part of the Penn Treebank 3 release (Marcus et al., 1993), the string in (1) is a variation 12-gram since *off* is a variation nucleus that in one corpus occurrence is tagged as a preposition (IN), while in another it is tagged as a particle (RP). Dickinson (2005) shows that examining those cases with identical local context (e.g., *ward off a*) results in an error detection precision of 92.8%; we will refer to these cases as *localized variation nuclei*.

(1) to ward **off** a hostile takeover attempt by two European shipping concerns

Why is this useful here? Not only are these contexts (i.e., trigrams) short and often recurring, but they seem to predict the usage of the same tag, at least when the annotation scheme is held constant. This is precisely the property we desire: a context which uniquely identifies a tag using one tagset means that the same decision could have been made when using another tagset.

Furthermore, these contexts are psychologically plausible contexts for category acquisition. Mintz (2002, 2003, 2006) shows that local word context predicts the same category for human category acquisition. Thus, both error detection and category acquisition point to a word immediately surrounded by identical words as a useful predictor of a single category. Indeed, Mintz (2003) points out that local context generally predicts the same basic class, even across different child-directed corpora.

The claim we are making is that the same context predicts the same category, and here we define a context as a localized variation nucleus. Mintz defines a context as a *frame* consisting only of the immediately surrounding words—i.e., not including the nucleus—and one could attempt to use such a context for cross-corpus comparison. However, being unsure of its precision for even one corpus, we use localized variation nuclei. For other languages, other contextual features might be more appropriate, such as affix information (cf., e.g., Tseng et al., 2005).

The claim is further that these contexts only predict the same *basic* category, such as verb; more fine-grained distinctions may or may not be present across corpora, and thus by predicting the same basic category, we can isolate spots where one corpus defines categories differently. In example (2), for instance, we see how the Brown corpus (see section 4.) uses the tag vb (verb) for both present tense (2a) and baseform (2b) instances. The Penn Treebank version of the Brown corpus, however, makes a distinction, between these two, as shown in (3).

(2) a. So you **know/vb** something of the classics
b. do you **know/vb** something I don't know?

(3) a. So you **know/VBP** something of the classics
b. do you **know/VB** something I don't know?

If this claim is true across corpora, then local contexts can be compared to spot important differences in tagset definitions. Thus, localized variation nuclei can be employed as our units of data.

3. Tagset comparison

However we use units of comparison (localized nuclei) to compare tagsets as a whole, we know that errors could cause spurious differences or highlight areas of difficulties, and so we first explore annotation error detection as a way to compare tagsets and then look at more specific differences across corpora.

3.1. Annotation error detection

While the variation n-gram method is highly accurate for the WSJ, there has been no thorough evaluation across corpora, to determine the effect of the tagset. Part of our purpose is to determine how comparable localized contexts are within and across corpora, and more fully evaluating the variation n-gram method gets at this question. Previously testing the method on the BNC-sampler (Leech, 1997) and finding much lower accuracy (52%) (Dickinson, 2005) did not take into account the differences between a news corpus (WSJ) and a balanced corpus (BNC-sampler). In other words, we do not know whether the difference in precision is more attributable to the tagset or to the differences in genre.

Thus, as a first method for comparing tagsets, we propose examining the output of the variation n-gram method, comparing the error detection precision and also the qualitative differences in tag variations. The more precise error detection is, the more local the tag decisions are; likewise, lower precision tends to indicate more non-local distinctions. This analysis will not only serve as a method of tagset comparison in its own right, but it will help us interpret the results of tagset discrepancies, in that we will know which variations are more likely errors.

3.2. Tagset discrepancies

Using the same units of comparison, we can more specifically examine disagreements in variation for the same localized nuclei. Discrepancies in how much variation there is for a given localized nucleus can indicate significant differences between the tagsets.

The types of tags which vary in one annotation scheme differ in another, and we can use this insight to develop a more automatic method of comparison. We can therefore identify those distinctions which vary in one corpus, but have little or no corresponding variation in the other. For example, as we will see in section 4., the tags preposition (IN) and particle (RP) vary a great deal in one corpus, but the corresponding situations in the other corpus are generally a single tag. The most problematic distinctions to account for in moving from tagset to tagset should be these exact discrepancies. Because this approach focuses on comparable contexts, it should turn up both problematic cases in the tagsets and differences in the locality of the tagsets.

In essence, tracking variation across corpora establishes points where tagsets do not map perfectly from one to another. When both annotation schemes license non-variation, it appears that one tag maps to another, and when both license variation, there is again a potentially straightforward mapping; it is when one has variation and the other does not that we find important discrepancies between the tagsets.

To calculate discrepancies, we investigate how much a variation between tags X and Y (denoted here as X/Y) maps to a single tag. As mentioned at the outset, this will indicate not only non-local tag differences, but differences in the difficulty of tagging for humans, as we will see in section 4.3.

Alternative calculations For qualitative evaluation, we will find that automatically generating these discrepancies is sufficient for turning up crucial differences between tagsets. One could explore more refined calculations, of course.

For example, one could factor out cases of corresponding variation, by using a difference metric. That is, one could count the number of variations featuring a distinction in one corpus and its corresponding variation or non-variation tag(s) in the other. Using these correspondences, one could then calculate how often a tag variation in one corpus corresponded to no variation in the other minus how often there was variation in the other corpus. For example, there are a total of 125 JJ/NN variations in one corpus which have no variation in the other versus 21 JJ/NN variations with corresponding variation, leading to a score of 104.

One could also explore all possible mappings between tagsets. In other words, one could compare how often JJ/NN maps to jj against how often it maps to nn (or perhaps some other set of tags).

Since we are really only interested in the discrepancies between the tagsets, we are simply examining the direct correspondences between variations and non-variations, as this is the most primary calculation. The alternative metrics are based upon these individual discrepancies. It is also important to point out that the use of comparable contexts readily supports all these possible calculations.

4. Evaluation

4.1. The Data

On the one hand, to emphasize the general nature of localized variation nuclei across texts, we want to run our experiments on a variety of texts, i.e., a balanced corpus. On the other hand, to control for differences between corpora and focus on differences in the tagset, we would at first like the text to be highly similar across annotations. The Brown corpus (Kucera and Francis, 1967) is thus an ideal test case: this balanced corpus contains its original annotation, but it was also re-annotated with a related, but simpler, tagset as part of the Penn Treebank (Marcus et al., 1993). We refer to these corpora as Brown and Brown-PTB, respectively.

Due to tokenization differences, Brown contains 1,161,192 tokens, and Brown-PTB has 1,170,811 tokens. In addition to slight textual differences, the corpora differ in tagset granularity, with Brown having 86 tags and Brown-PTB 45. While a full analysis of these two corpora will show us much about the method, using two tagsets can only begin to give us an indication of how our methods can be used for comparison. For a more thorough analysis, we additionally explore a third corpus—which happens also to be based on the Brown data—for the tagset discrepancy calculation (section 4.3.). The SUSANNE corpus (Sampson, 1995) has approximately 151,600 tokens and provides a much more fine-grained annotation scheme, with 424 lexical categories.[1] Comparing it to the other two tagsets will provide insights into whether and how fine-grained distinctions make a difference in context, i.e., whether they are actually more difficult to disambiguate or not.[2]

4.2. Annotation error detection

Running the variation n-gram method on the first two corpora results in 1605 localized variation nuclei for Brown and 1809 for Brown-PTB. The amount of variation is thus roughly comparable, but they differ widely in their error detection precision: from these sets, we sampled 100 cases of each and marked for each variation whether it pointed to an error in POS annotation. For Brown, we find 42% precision and for Brown-PTB, 84%, indicating that legitimate non-locality is a greater problem in Brown and that errors (and unclear distinctions) are more problematic for Brown-PTB. We thus confirm that local context has a critically different impact on different tagsets: as there are fewer legitimate ambiguities in Brown-PTB, disambiguating words in Brown-PTB appears to be, in some sense, slightly easier than in Brown.

Perhaps as important as the absolute difference in error detection precision are the reasons for the differences, stemming from both the general quality of the annotation effort and the distinctions made in the tagsets. Brown, for example, uses function tags to indicate, e.g., that the word is being used in a headline (hl). This is clearly a non-local use, and 32 variations involved the hl tag. Removing these gives 62% precision (42/68), an improvement, but still considerably lower than for Brown-PTB.

The correctly ambiguous cases in both corpora involve distinctions which often require non-local information, e.g., VB/VBP (baseform verb/present tense verb) in Brown-PTB and cs/in (subordinating conjunction/preposition) in Brown (see also the discussion in section 4.3. below).

There were also several cases which were unclear, such as the distinction between adjective (JJ) and noun (NN) in Brown-PTB, which is perhaps not adequately explicated for compounds in the guidelines (Santorini, 1990), leading to variation as in (4).

(4) a. ... through most of the **liquid/JJ** helium .
b. ... to isolate the tube ... from the **liquid/NN** helium surface

Finally, many of the variations reveal constructions which are difficult to analyze linguistically, most involving comparative constructions, such as *and as far as X*, where the first *as* erroneously varies between adverb (RB) and preposition (IN) in Brown-PTB, as in (5).

(5) a. ...or fine points treated singly, and **as/RB** far as possible impersonally.
b. The Class had entries from as far west as Wisconsin and **as/IN** far south as Kentucky.

[1] This token count does not include *break* and *ghost* tokens.

[2] Brants (1995), for example, shows that tagsets can be reduced without losing information.

In *not nearly as X*, *nearly* correctly varies between qualifier (ql) and adverb (rb) in Brown—a distinction collapsed into RB in Brown-PTB—as shown in (6). Determining these cases not only often requires more context to disambiguate, but some knowledge of linguistic theory (see, e.g., Pollard and Sag, 1994, sec. 9.4).

(6) a. not **nearly/rb** as complex/jj
b. not **nearly/ql** as much/ap

Manually examining the error detection output thus uncovers tag distinctions which are problematic in one way or another in one tagset. We still need to pinpoint the exact differences between tagsets, and we thus turn to the tagset discrepancy calculations.

4.3. Tagset discrepancies

Brown and Brown-PTB The highest-ranking correspondences between the Brown and Brown-PTB tagsets are shown in figure 1. We can notice a variety of things from this chart.

PTB	Brown	Count
VB/VBP	vb	124
IN/RP	rp	105
IN	cs/in	98
IN/RB	rp	86
JJ/NN	jj	71
IN	in/in-hl	66
DT	at/at-hl	65
NN	nn/nn-hl	56
.	./.-hl	56
IN	in/rp	51
DT	at/at-tl	45
DT/PDT	abn	44
JJ/VBN	vbn	41
RB	ql/rb	40
VBD/VBN	vbn	39
VB/VBP	hv	37
JJ/NN	nn	37
NN/NNP	nn-tl	35
NNP	np/np-tl	34
NNP/NNPS	nns-tl	32
JJR/RBR	ap	30

Figure 1: Correspondences between Brown & Brown-PTB

First of all, as mentioned above, the clearly non-local -hl tag adds many spurious variations, which have a clear correspondent in the other tagset.
Secondly, the highest scores in both directions (VB/VBP $\mapsto$ vb and cs/in $\mapsto$ IN) reflect tags which are combined in the other tagset. The non-local distinction cs/in in Brown never has a corresponding variant in Brown-PTB because these tags were conflated into IN; likewise, VB/VBP in Brown-PTB is conflated into vb in Brown.
Thirdly, aside from cases of variation with function tags, there is much more variation for Brown-PTB than for Brown. This high variation results from at least two major factors.

Firstly, clear differences in the tagsets can lead to variation in one corpus but not the other. In Brown, for example, verbs ending in *-ed* are tagged vbn (past participle), even when acting as a noun modifier. However, as can be seen in (7), in Brown-PTB these lexical forms may also be tagged as JJ. This new distinction between JJ and VBN in the tagset can make disambiguation more difficult.

(7) a. ...impartial/jj and **standardized/vbn** procedure
b. ...impartial/JJ and **standardized/JJ** procedure

Secondly, we also find that a number of annotation errors stem from these tagset differences, as the introduction of a new distinction may be one which is difficult to make, not just for automatic taggers, but also for humans. With this distinction between adjectives and past participles, we thus find clear examples of inconsistency in the annotation, as in (8), where both should be JJ.

(8) a. You would be **surprised/VBN** how ...
b. he would not be **surprised/JJ** if ...

Other examples, like (9), are less clear as to whether they are an error or simply difficult to disambiguate. Thus, for all these reasons, it is no surprise that we find 41 cases of variation of JJ/VBN in Brown-PTB, corresponding to a single tag, vbn, in Brown.

(9) a. After well broken/JJ and **equipped/JJ** with 12-oz. shoes on behind, bare-footed in front, she would trot...
b. The Royal Lao Army, on the other hand, was paid/VBN and **equipped/VBN** with American funds.

To take another example, IN, RB (adverb), RP (particle) in Brown-PTB are defined differently than their counterparts in Brown, and so we find a high number of correspondences for the three pairs IN/RP (105), IN/RB (86), and RB/RP (26) all of which most frequently correspond to the single tag rp in Brown.[3] This distinction in Brown-PTB is difficult to disambiguate and fraught with errors; as it turns out, the notion of adverb and particle is more lexicalized in Brown, which leads to such correspondences. As it states in the Brown manual (Francis and Kucera, 1979),

> It was decided instead to consider this a syntactic and semantic rather than a taxonomic problem, and to give the "portmanteau" tag RP (for "adverb or particle") to the ten words *about*, *across*, *down*, *in*, *off*, *on*, *out*, *over*, *through*, and *up*, except when they are functioning as prepositions, when they receive the normal preposition tag IN.

Thus, we often find non-variation for these words, and without such a strict lexical principle, there is (often erroneous) corresponding variation in Brown-PTB, as in (10).

(10) a. ...the back pressure of the manometer was built **up/RP** from the material fed from between the blocks...

[3] Additionally, variation between all three tags, IN/RB/RP, corresponds to rp 24 times.

b. The fixed wooden scaffold was removed, and so as to reach all the frieze, one of pipe, on wheels, built **up/IN** from the floor.

We have turned up cases where the tagsets do not align, in ways that lead to dramatic differences in difficulty. This is crucial in understanding tag correspondences (e.g., Manning and Schütze, 1999, table 4.5) and their definitional differences, and also in understanding that a tagger optimized on one tagset will not necessarily transfer to another. For example, in order to disambiguate adverbs, prepositions, and particles, Toutanova and Manning (2000) use a feature "adding information on verbs' preferences to take specific words as particles, or adverbs, or prepositions." This feature, while useful for a tagset such as Brown-PTB's, is less useful for Brown since particle (rp) is often a lexical property, and there is much less confusion.

Brown and SUSANNE Turning to a comparison of the Brown and SUSANNE corpora, we know that SUSANNE has a much finer-grained tagset. However, we see in figure 2 that there is more variation in Brown than in SUSANNE for the same local contexts. In other words, granularity is not a good indicator of difficulty of tagging (cf, e.g., Voutilainen and Järvinen, 1995).

Brown	SUSANNE	Count
cs/in	CSN	19
at/at-tl	AT	15
at/at-hl	AT	15
vbd/vbn	VVDv	13
in/rp	RP	13
)/)-hl	)	13
jj/rb	JJ	11
ql/rb	RR	8
pp$/ppo	APPGf	8
ap/rbr	DAR	8
cs/in	ICSt	7
vbd/vbn	VVNv	5
jj/nn	JJ	5
in	II33/IO	5
(/(-hl	(	5

Figure 2: Correspondences between Brown & SUSANNE

Like the previous comparison between Brown and Brown-PTB, with Brown and SUSANNE we find that using the correspondences between tags highlights tagset distinctions, and again pinpoints errors in one tagset versus the other. Lexical distinctions between the tagsets play an important role, as can be seen with the first item on the list, cs/in versus CSN. CSN is used for all instances of the word *than* in SUSANNE, while Brown makes the grammatical distinction between subordinating conjunction (cs) and preposition (in).

As mentioned, SUSANNE makes more distinctions than Brown, but these do not show up as problems in context, meaning that these distinctions are more clearly defined, local, and/or infrequent. On the other hand, consider the case of ql/rb (qualifiers/adverbs). This disinction in Brown is more difficult to disambiguate, as it requires non-local context (cf. (6)). This confirms our hypothesis that the type of variations we find are ones which break the assumption of needing only local context to disambiguate.

The distinction between possesive pronouns (pp$) and accusative pronouns (ppo) in Brown is also found in SUSANNE, and the fact that the variation does not occur in both leads us to discover that this is an error in Brown. In this case it is due to an error annotating *her* as accusative instead of possessive. As examples like this show, highlighting cases of variation in one corpus without variation in another corpus can actually be a way to improve error detection precision: we are more confident that something is amiss when the other corpus has no variation.

Note that, for counts of five or higher, the only variation in SUSANNE which does not have variation in Brown is II33/IO, but this is easily disambiguated in context because it is a ditto tag that is tagged in conjunction with the previous tags II31 and II32.[4]

Brown-PTB and SUSANNE Examining the differences in Brown-PTB and SUSANNE, as shown in figure 3, turns up many similar cases to the discrepancies between Brown and SUSANNE above, likely due to their shared history. Again, the coarser-grained tagset winds up having more variation in context. This is also likely related to annotation quality (cf. the comparison between the WSJ and SUSANNE in Dickinson, 2005)

PTB	SUSANNE	Count
JJ/NN	JJ	25
IN/RP	RP	23
IN/RB	RP	21
DT/PDT	DBa	17
JJR/RBR	DAR	15
JJ/NNP	JJ	13
VBD/VBN	VVNv	12
IN/RB/RP	RP	8
WDT/WP	DDQ	6
VBD/VBN	VVDv	6
JJS/RBS	DAT	6
DT/RB	ATn	6
PRP/PRP$	APPGf	5
JJ/PDT	DAz	5
IN/RB	II	5
DT/JJ/PDT	DAz	5
IN	II33/IO	5

Figure 3: Correspondences between Brown-PTB & SUSANNE

Even though both tagsets make use of adjective (JJ) and particle (RP) tags, they seem to be more easily annotated in SUSANNE, i.e., with less variation. Given the results in section 4.2., it is likely that a good portion of these variations are errors, indicating the difficulty with which annotators had making the distinctions.

[4]The ditto tag is used to tag words in a sequence, like "so as to", with the part-of-speech of the whole sequence as opposed to the individual words.

5. Summary and Outlook

We have developed a simple method to easily compare tagsets, in order to both understand the linguistic distinctions employed in a tagset and the impact of differing categories on the technology using them. This is important in porting technology from one tagset to another and in providing feedback on which parts of an annotation scheme are most difficult. By using a simple notion of local context which easily transfers across corpora in the same language, tagset discrepancies can be highlighted which point out the major differences in category definitions.

The method was based on a notion of local context which has been shown to be useful in human category acquisition. In comparing categories across corpora, we are beginning to employ local context to identify areas of identical disambiguation. Thus, future work could more fully explore the information needed for disambiguation across corpora and extend into category induction. One crucial step in this work will be to compare corpora of different genres and different languages, in order to see how general localized variation nuclei are in their prediction of the same basic category.

In parallel with the development of error detection methods (Dickinson, 2005), the methods explored here can also potentially be extended to more complex forms of annotation, where there is an even greater need for comparison, such as syntactic annotation (cf. Rehbein and van Genabith, 2007).

References

Babarczy, Anna, John Carroll, and Geoffrey Sampson (2006). Definitional, personal, and mechanical constraints on part of speech annotation performance. *Journal of Natural Language Engineering* 12, 77–90.

Brants, Thorsten (1995). Tagset Reduction Without Information Loss. In *Proceedings of ACL-95*. Cambridge, MA.

Brants, Thorsten (1997). Internal and External Tagsets in Part-of-Speech Tagging. In *Proceedings of Eurospeech*. Rhodes, Greece.

Déjean, Hervé (2000). How to Evaluate and Compare Tagsets? A Proposal. In *Proceedings of LREC-00*. Athens.

Dickinson, Markus (2005). Error detection and correction in annotated corpora. Ph.D. thesis, The Ohio State University.

Dickinson, Markus and W. Detmar Meurers (2003). Detecting Errors in Part-of-Speech Annotation. In *Proceedings of EACL-03*. Budapest, Hungary, pp. 107–114.

Francis, W. N. and H. Kucera (1979). *Brown Corpus Manual*. Tech. rep., Department of Linguistics, Brown University, Providence, Rhode Island, US.

Kucera, Henry and W. Nelson Francis (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Leech, Geoffrey (1997). *A Brief Users' Guide to the Grammatical Tagging of the British National Corpus*. UCREL, Lancaster University.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

Mintz, Toben H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* 30, 678–686.

Mintz, Toben H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90, 91–117.

Mintz, Toben H. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek and R. M. Golinkoff (eds.), *Action Meets Word: How Children Learn Verbs*, New York: Oxford University Press, pp. 31–63.

Pollard, Carl and Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rehbein, Ines and Josef van Genabith (2007). Treebank Annotation Schemes and Parser Evaluation for German. In *Proceedings of EMNLP-CoNLL-07*. pp. 630–639.

Sampson, Geoffrey (1995). *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Santorini, Beatrice (1990). *Part-Of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd printing)*. Tech. Rep. MS-CIS-90-47, The University of Pennsylvania, Philadelphia, PA.

Toutanova, Kristina and Christopher D. Manning (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of EMNLP/VLC-2000*. Hong Kong.

Tseng, Huihsin, Daniel Jurafsky and Christopher Manning (2005). Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Voutilainen, Atro and Timo Järvinen (1995). Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of EACL-95*. Dublin, Ireland, pp. 210–214.