Categorizing Local Contexts as a Step in Grammatical Category Induction

Markus Dickinson
Indiana University
Bloomington, IN USA
md7@indiana.edu

Charles Jochim
Indiana University
Bloomington, IN USA
cajochim@indiana.edu

Abstract

Building on the use of local contexts, or frames, for human category acquisition, we explore the treatment of contexts as categories. This allows us to examine and evaluate the categorical properties that local unsupervised methods can distinguish and their relationship to corpus POS tags. From there, we use lexical information to combine contexts in a way which preserves the intended category, providing a platform for grammatical category induction.

1 Introduction and Motivation

In human category acquisition, the immediate local context of a word has proven to be a reliable indicator of its grammatical category, or part of speech (e.g., Mintz, 2002, 2003; Redington et al., 1998). Likewise, category induction techniques cluster word types together (e.g., Clark, 2003; Schütze, 1995), using similar information, i.e., distributions of local context information. These methods are successful and useful (e.g. Koo et al., 2008), but in both cases it is not always clear whether errors in lexical classification are due to a problem in the induction algorithm or in what contexts count as identifying the same category (cf. Dickinson, 2008). The question we ask, then, is: what role does the context *on its own* play in defining a grammatical category? Specifically, when do two contexts identify the same category?

Many category induction experiments start by trying to categorize words, and Parisien et al. (2008) categorize *word usages*, a combination of a word and its context. But to isolate the effect the context has on the word, we take the approach of categorizing contexts as a first step towards clustering words. By separating out contexts for word clustering, we can begin to speak of better disambiguation models as a foundation for induction. We aim in this paper to thoroughly investigate what category properties contexts can or cannot distinguish by themselves.

With this approach, we are able to more thoroughly examine the categories used for evaluation. Evaluation of induction methods is difficult, due to the variety of corpora and tagsets in existence (see discussion in Clark, 2003) and the variety of potential purposes for induced categories (e.g., Koo et al., 2008; Miller et al., 2004). Yet improving the evaluation of category induction is vital, as evaluation does not match up well with grammar induction evaluation (Headden III et al., 2008). For many evaluations, POS tags have been mapped to a smaller tagset (e.g., Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), but there have been few criteria for evaluating the quality of these mappings. By isolating contexts, we can investigate how each mapping affects the accuracy of a method and the lexicon.

Using corpus annotation also allows us to explore the relation between induced categories and computationally or theoretically-relevant categories (e.g., Elworthy, 1995). While human category acquisition results successfully divide a lexicon into categories, these categories are not necessarily ones which are appropriate for many computational purposes or match theoretical syntactic analysis. This work can also serve as a platform to help drive the design of new tagsets, or refinement of old ones, by outlining which types of categories are or are not applicable for category induction.

After discussing some preliminary issues in section 2, in section 3 we examine to what extent contexts by themselves can distinguish different category properties and how this affects evaluation. Namely, we propose that corpus tagsets should be clear about identifying syntactic/distributional properties and about how tagset mappings for evaluation should outline how much information

Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition, pages 34–41,
Athens, Greece, 31 March 2009. ©2009 Association for Computational Linguistics

is lost by mapping. In section 4, in more preliminary work, we add lexical information to contexts, in order to merge them together and see which still identify the same category.

2 Preliminaries

2.1 Background

Research on language acquisition has addressed how humans learn categories of words, and we use this as a starting point. Mintz (2002) shows that local context, in the form of a *frame* of two words surrounding a target word, leads to the target's categorization in adults, and Mintz (2003) shows that frequent frames supply category information in child language corpora. A frame is not decomposed into its left and right sides (cf., e.g., Redington et al., 1998; Clark, 2003; Schütze, 1995), but is taken as their joint occurrence (Mintz, 2003).[1]

For category acquisition, *frequent frames* are used, those with a frequency above a certain threshold. These predict category membership, as the set of words appearing in a given frame should represent a single category. The frequent frame *you* __ *it*, for example, largely identifies verbs, as shown in (1), taken from child-directed speech in the CHILDES database (MacWhinney, 2000). For frequent frames in six subcorpora of CHILDES, Mintz (2003) obtains both high type and token accuracy in categorizing words.

(1) a. you put it
    b. you see it

The categories do not reflect fine-grained linguistic distinctions, though, nor do they fully account for ambiguous words. Indeed, accuracies slightly degrade when moving from "Standard Labeling"[2] to the more fine-grained "Expanded Labeling,"[3] from .98 to .91 in token accuracy and from .93 to .91 in type accuracy. In scaling the method beyond child-directed speech, it would be beneficial to use annotated data, which allows for ambiguity and distinguishes a word's category across corpus instances. Furthermore, even though many frames identify the same category, the method does not thoroughly specify how to relate them.

It has been recognized for some time that wider contexts result in better induction models (e.g., Parisien et al., 2008; Redington et al., 1998), but many linguistic distinctions rely on lexical information that cannot be inferred from additional context (Dickinson, 2008), so focusing on short contexts can provide many insights. The use of frames allows for frequent recurrent contexts and a way to investigate corpus categories, or POS tags (cf., e.g., Dickinson and Jochim, 2008). An added benefit of starting with this method is that it can be converted to a model of online acquisition (Wang and Mintz, 2007). For this paper, however, we only investigate the type of information input into the model.

2.2 Some definitions

Frequency The core idea of using frames is that words used in the same context are associated with each other, and the more often these contexts occur, the more confidence we have that the frame indicates a category. Setting a threshold to obtain the 45 most frequent frames in each subcorpus (about 80,000 words on average), (Mintz, 2003) allows a frame to occur often enough to be meaningful and have a variety of target words in the frame.

To determine what category properties frames pinpoint (section 3), we use two thresholds to define *frequent*. Singly occurring frames cannot provide any information about groupings of words, so we first consider frames that occur more than once. This gives a large number of frames, covering much of the corpus (about 970,000 tokens), but frames with few instances have very little information. For the other threshold, frequent frames are those which have a frequency of 200, about 0.03% of the total number of frames in the corpus. One could explore more thresholds, but for comparing tagset mappings, these provide a good picture. The higher threshold is appropriate for combining contexts (section 4), as we need more information to tell whether two frames behave similarly.

Accuracy To evaluate, we need a measure of the accuracy of each frame. Mintz (2003) and Redington et al. (1998) calculate accuracy by counting all pairs of words (types or tokens) that are from the same category, divided by all possible pairs of words in a grouping. This captures the idea that each word should have the same category as every

[1] This use of *frame* is different than that used for subcategorization frames, which are also used to induce word classes (e.g., Korhonen et al., 2003).

[2] Categories = noun, verb, adjective, preposition, adverb, determiner, wh-word, *not*, conjunction, and interjection.

[3] Nouns split into nouns and pronouns; verbs split into verbs, auxiliaries, and copula

other word in its category set.

Viewing the task as disambiguating contexts (see section 3), however, this measurement does not seem to adequately represent cases with a majority label. For example, if three words have the tag X and one Y, pairwise comparison results in an accuracy of 50%, even though X is dominant. To account for this, we measure the precision of the most frequent category instances among all instances, e.g., 75% for the above example (cf. the notion of *purity* in Manning et al., 2008). Additionally, we only use measurements of token precision. Token precision naturally handles ambiguous words and is easy to calculate in a POS-annotated corpus.

3 Categories in local contexts

In automatic category induction, a category is often treated as a set, or cluster, of words (Clark, 2003; Schütze, 1995), and category ambiguity is represented by the fact that words can appear in more than one set. Relatedly, one can cluster *word usages*, a combination of a word and its context (Parisien et al., 2008). An erroneous classification occurs when a word is in an incorrect set, and one source of error is when the contexts being treated as indicative of the same category are actually ambiguous. For example, in a bigram model, the context *be* __ identifies nouns, adjectives, and verbs, among others.

Viewed in this way, it is important to gauge the precision of contexts for distinguishing a category (cf. also Dickinson, 2008). In other words, how often does the same context identify the same category? And how fine-grained is the category that the context distinguishes? To test whether a frame defines a single category in non-child-directed speech, we focus on which categorical properties frames define, and for this we use a POS-annotated corpus. Due to its popularity for unsupervised POS induction research (e.g., Goldberg et al., 2008; Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008) and its often-used tagset, for our initial research, we use the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1993), with 36 tags (plus 9 punctuation tags), and we use sections 00-18, leaving held-out data for future experiments.[4]

Defining frequent frames as those occurring at least 200 times, we find 79.5% token precision. Additionally, we have 99 frames, identifying 14 types of categories as the majority tag (common noun (NN) being the most prevalent (37 frames)). For a threshold of 2, we have 77.3% precision for 67,721 frames and 35 categories.[5] With precision below 80%, we observe that frames are not fully able to disambiguate these corpus categories.

3.1 Frame-defined categories

These corpus categories, however, are composed of a variety of morphological and syntactic features, the exact nature of which varies from tagset to tagset. By merging different tags, we can factor out different types of morphological and syntactic properties to determine which ones are more or less easily identified by frames. Accuracy will of course improve by merging tags; what is important is for which mappings it improves.

We start with basic categories, akin to those in Mintz (2003). Despite the differences among tagsets, these basic categories are common, and merging POS tags into basic categories can show that differences in accuracy have more to do with stricter category labels than language type. We merged tags to create basic categories, as in table 1 (adapted from Hepple and van Genabith (2000); see appendix A for descriptions).[6]

Category	Corpus tags
Determiner	DT, PDT, PRP$
Adjective	JJ, JJR, JJS
Noun	NN, NNS, PRP, NNP, NNPS
Adverb	RB, RBR, RBS
Verb	MD, VB, VBD, VBG, VBN, VBP, VBZ
Wh-Det.	WDT, WP$

Table 1: Tag mappings into basic categories

These broader categories result in the accuracies in table 2, and we also record accuracies for the similar PTB-17 tagset used in a variety of unsupervised tagging experiments (Smith and Eisner, 2005), which mainly differs by treating VBG and VBN uniquely. With token precision around 90%, it seems that frame-based disambiguation is generally identifying basic categories, though with less

[4] Even if we wanted child-directed speech, the CHILDES database (MacWhinney, 2000) uses coarse POS tags.

[5] LS (List item marker) is not identified; UH (interjection) appears in one repeating frame, and SYM (symbol) in two.

[6] The 13 other linguistic tags were not merged, i.e., CC, CD, EX, FW, IN, LS, POS, RP, SYM, TO, UH, WP, WRB.

accuracy than in Mintz (2003).

	≥ 2	≥ 200
Orig.	77.3%	79.5%
Merged	85.9%	91.0%
PTB-17	85.1%	89.7%

Table 2: Effect of mappings on precision

But which properties of the tagset do the frame contexts accurately capture and which do they not? To get at this question, we explore linguistically-motivated mappings between the original tagset and the fully-merged tagset in table 1. Given the predominance of verbs and nouns, we focus on distinguishing linguistic properties within these categories. For example, simply by merging nouns and leaving all other original tags unchanged, we move from 79.5% token precision to 88.4% (for the threshold of 200).

Leaving all other mappings as in table 1, we merge nouns and verbs along two dimensions: their common syntactic properties or their common morphological properties. Ideally, we prefer frames to pick out syntactic properties, since morphological properties can assumedly be determined from word-internal properties (see Clark, 2003; Christiansen and Monaghan, 2006).

Specifically, we can merge nouns by *noun type* (PRP [pronoun], NN/NNS [common noun], NNP/NNPS [proper noun]) or by *noun form*, in this case based on grammatical number (PRP [pronoun], NN/NNP [singular noun], NNS/NNPS [plural noun]). We can merge verbs by *finiteness* (MD [modal], VBP/VBZ/VBD [finite verb], VB/VBG/VBN [nonfinite verb]) or by *verb form* (MD [modal], VB/VBP [base], VBD/VBN [-*ed*], VBG [-*ing*], VBZ [-*s*]). In the latter case, verbs with consistently similar forms are grouped—e.g., *see* can be a baseform (VB) or a present tense verb (VBP).

The results are given in tables 3 and 4. We find that merging verbs by finiteness and nouns by noun type results in higher precision. This confirms that contexts can better distinguish syntactic, but not necessarily morphological, properties. As we will see in the next section, this mapping also maintains distinctions in the lexicon. Such use of local contexts, along with tag merging, can be used to evaluate tagsets which claim to be distributional (see, e.g., Dickinson and Jochim, 2008).

It should be noted that we have only explored category mappings which merge tags, ignoring possible splits. While splitting a tag like TO (*to*) into prepositional and infinitival uses would be ideal, we do not have the information automatically available. We are thus limited in our evaluation by what the tagset offers. Some tag splits can be automatically recovered (e.g., splitting PRP based on properties such as person), but if it is automatically recoverable from the lexicon, we do not necessarily need context to identify it, an idea we turn to in the next section.

	Noun type	Noun form
Finiteness	**82.9%**	81.2%
Verb form	81.2%	79.5%

Table 3: Mapping precision (freq. ≥ 2)

	Noun type	Noun form
Finiteness	**86.4%**	85.3%
Verb form	84.5%	83.4%

Table 4: Mapping precision (freq. ≥ 200)

3.2 Evaluating tagset mappings

Some of the category distinctions made by frames are more or less important for the context to make. For example, it is detrimental if we conflate VB and VBP because this is a prominent ambiguity for many words (e.g., *see*). On the other hand, there are no words which can be both VBP (e.g., *see*) and VBZ (e.g., *sees*). Ideally, induction methods would be able to distinguish all these cases—just as they often make distinctions beyond what is in a tagset—but there are differences in how problematic the mappings are. If we group VB and VBP into one tag, there is no way to recover that distinction; for VBP and VBZ, there are at least different words which inherently take the different tags.

Thus, a mapping is preferred which does not conflate tags that vary for individual words. To calculate this, we compare the original lexicon with a mapped lexicon and count the number of words which lose a distinction. Consider the words *accept* and *accepts*: *accept* varies between VB and VBP; *accepts* is only VBZ. When we map tags based on verb form, we count 1 for *accept*, as VB and VBP are now one tag (Verb). When we map verbs based on finiteness, we count 0 for these two words, as *accept* still has two tags (V-nonfin, V-fin) and *accepts* has one tag (V-fin).

We evaluate our mappings in table 5 by enumerating the number of word types whose distinctions are lost by a particular mapping (out of 44,520 word types); we also repeat the token precision values for comparison. Perhaps unsurprisingly, grouping words based on form results in high confusability (cf. the discussion of *see* in section 3.1). On the other hand, merging nouns by type and verbs by finiteness results in something of a balance between precision and non-confusability. It is thus these types of categorizations which we can reasonably expect induction models to capture.

	Lost	Precision	
Mapping	tags	≥ 2	≥ 200
All mappings	3003	85.9%	91.0%
PTB-17	2038	85.1%	89.7%
N. form/V. form	2699	79.5%	83.4%
N. type/V. form	2148	81.2%	84.5%
N. form/Finite	951	81.2%	85.3%
N. type/Finite	**399**	**82.9%**	**86.4%**
No mappings	0	77.3%	79.5%

Table 5: Confusable word types

For induction evaluation, in addition to an accuracy metric, a metric such as the one we have just proposed is important to gauge how much corpus annotation information is lost when performing tagset mappings. For example, the PTB-17 mapping (Smith and Eisner, 2005) is commonly used for evaluating category induction (Goldwater and Griffiths, 2007; Toutanova and Johnson, 2008), yet it loses distinctions for 2038 words.

We could also define mappings which lose no distinctions in the lexicon. Initial experiments show that this allows no merging of nouns, and that the resulting precision is only minimally better than no mapping at all. We should also note that the number of confusable words may be too high, given errors in the lexicon (cf. Dickinson, 2008). For example, removing tags occurring less than 10% of the time for a word results in only 305 confusable words for the Noun type/Finiteness (NF) mapping and 1575 for PTB-17.

4 Combining contexts

We have narrowly focused on identical contexts, or frames, for identifying categories, but this could leave us with as many categories as frames (67,721 for ≥ 2, 99 for ≥ 200, instead of 35 and 30). We need to reduce the number of categories without inappropriately merging them (cf. the notion of "completeness" in Mintz, 2003; Christiansen and Monaghan, 2006). Thus far, we have not utilized a frame's target words; we turn to these now, in order to better gauge the effectiveness of frames for identifying categories. Although the work is somewhat preliminary, our goal is to continue to investigate when contexts identify the same category. This merging of contexts is different than clustering words (e.g., Clark, 2000; Brown et al., 1992), but is applicable, as word clustering relies on knowing which contexts identify the same category.

4.1 Word-based combination

On their own, frames at best distinguish only very broad categorical properties. This is perhaps unsurprising, as the finer-grained distinctions in corpora seem to be based on lexical properties more than on additional context (see, e.g., Dickinson, 2008). If we want to combine contexts in a way which maps to corpus tagsets, then, we need to examine the target words. It is likely that two sets share the same tag if they contain the same words (cf. overlap in Mintz, 2003). In fact, the more a frame's word set overlaps with another's word set, the more likely it is unambiguous in the first place, as the other set provides corroborating evidence. Therefore, we use overlap of frames' word sets as a criterion to combine them.

This allows us to combine frames which do not share context words. For example, in (2) we find frames identifying baseform verbs (VB) (2a) and frames identifying cardinal numbers (CD) (2b), despite having a variety of context words. Their target word sets, however, are sufficiently similar.

(2) a. will __ to, will __ the, to __ the, to __ up,
would __ the, to __ their, n't __ the,
to __ a, to __ its, to __ that, to __ to

b. or __ cents, $ __ million, rose __ %,
a __ %, about __ %, to __ %, $ __ a,
$ __ billion

By viewing frames as categories, in the future we could also investigate splitting categories, based on subsets of words, morphological/phonological cues (e.g., Christiansen and Monaghan, 2006), or on additional context words, better handling frames that are ambiguous.

Calculating overlap We merge frames whose word sets overlap, using a simple weighted frequency distance metric. We define sufficient overlap as the case where a given percent of the words in one frame's word set are found in the other's word set. We define this test in either direction, as smaller sets can be a subset of a larger set. For example, the frames *the __ on* (224 tokens) and *the __ of* (4304 tokens) have an overlap of 78 tokens; overlap here is 34.8% (78/224). While we could use a more sophisticated form of clustering (see, e.g., Manning et al., 2008), this will help determine the viability of this general approach.

Of course, two sets may share a category with relatively few shared words, and so we transitively combine sets of contexts. If the overlap of frames A and B meet our overlap criterion and the overlap of frames A and C also meet the criterion, then all three sets are merged, even if B and C have only a small amount of overlap.[7]

Using the threshold of 200, we test criteria of 30%, 40%, and 50% overlap and consider the frames' overlap calculated as a percentage of word types or as a percentage of word tokens. For example, if a word type occurs 10 times in one word set and 20 in the other, the overlap of types is 1, and the overlap of tokens is 10. Token overlap better captures similarities in distributions of words.

4.2 Evaluation

Table 6 shows the number of categories for the 30%, 40%, and 50% type-based (TyB) and token-based (ToB) overlap criteria for merging. As we can see, the overlap based on tokens in word sets results in more categories, i.e., fewer merges.

%	TyB	ToB
50%	59	75
40%	42	64
30%	27	50

Table 6: Number of categories by condition

The precision of each of these criteria is given in table 7, evaluating on both the original tagset and the noun type/finiteness (*NF*) mapping. We can see that the token-based overlap is consistently more accurate than type-based overlap, and there is virtually no drop in precision for any of the token-based conditions.[8] Thus, for the rest of the evaluation, we use only the token-based overlap.

%	Tags	Frames	TyB	ToB
50%	Orig.	79.5%	76.4%	79.5%
	NF	86.4%	82.8%	86.4%
40%	Orig.	79.5%	75.7%	79.3%
	NF	86.4%	81.8%	86.1%
30%	Orig.	79.5%	74.7%	79.1%
	NF	86.4%	81.7%	86.1%

Table 7: Precision of merged frames

We mentioned that if frame word sets overlap, the less ambiguous their category should be. We check this by looking at the difference between merged and unmerged frames, as shown in table 8. The number of categories are also given in parentheses; for example, for 30% overlap, 41 frames are unmerged, and the remaining 58 make up 9 categories. These results confirm for this data that frames which are merged have a higher precision.

	Merged	Unmerged	Overall
50%	93.4% (7)	79.9% (68)	86.4% (75)
40%	89.7% (10)	81.1% (54)	86.1% (64)
30%	89.7% (9)	77.4% (41)	86.1% (50)

Table 8: Precision of merged & unmerged frames for NF mapping (with number of categories)

But are we only merging a select, small set of words? To gauge this, we measure how much of the corpus is categorized by the 99 most frequent frames. Namely, 46,874 tokens occur as targets in our threshold of 99 frequent frames out of 663,608 target tokens in the entire corpus,[9] a recall of 7.1%. Table 9 shows some recall figures for the frequent frames. There are 9621 word types in the set of target words for the 99 frequent frames, which is 27.2% of the target lexicon. Crucially, though, these 9621 are realized as 523,662 target tokens in the corpus, or 78.9%. The words categorized by the frequent frames extend to a large portion of the corpus (cf. also Mintz, 2003).

	Tokens	Types	Coverage
Merged (30%)	5.0%	20.0%	61.5%
Unmerged (30%)	2.0%	11.5%	65.9%
Total Overlap	7.1%	27.2%	78.9%

Table 9: Recall of frames

[7] We currently do not consider overlap of already merged sets, e.g., between A+B and C.

[8] Experiments at 20% show a noticeable drop in precision.

[9] Because we remove frames which contain punctuation, the set of target tokens is a subset of all words in the corpus.

4.2.1 Qualitative analysis

To better analyze what is happening for future work, we look more closely at 30% overlap. Of the 58 frames merged into 9 categories, 54 of them have the same majority tag after merging. The four frames which get merged into a different category are worth investigating, to see the method's limitations and potential for improvement.

Of the four frames which lose their majority tag after merging, two can be ignored when mapping to the NF tags. The frame *it* __ *the* with majority tag VBZ becomes VBD when merged, but both are V-fin. Likewise, *n't* __ *to* changes from VB to VBN, both cases of V-nonfin. The third case reveals an evaluation problem with the original tagset: the frames *million* __ *$* (IN) and *%* __ *$* (TO) are merged into a category labeled TO. The tag TO is for the word *to* and is not split into prepositional and infinitival uses. Corpus categories such as these, which overlap in their definitions yet cannot be merged (due to their non-overlapping uses), are particularly problematic for evaluation.

The final case which does not properly merge is the most serious. The frame *is* __ *the* (37% of tokens as preposition (IN)) merges with *is* __ *a* (41% of tokens as VBG); the merged VBG category has an precision of 34%. The distribution of tags is relatively similar, the highest percentages being for IN and VBG in both. This highlights the point made earlier, that more information is needed, to split the word sets.

4.2.2 TIGER Corpus

To better evaluate frequent frames for determining categories, we also test them on the German TIGER corpus (Brants et al., 2002), version 2, to see how the method handles data with freer word order and more morphological complexity. We use the training data, with the data split as in Dubey (2004). The frequency threshold for the WSJ (0.03% of all frames) leaves us with only 60 frames in the TIGER corpus, and 51 of these frames have a majority tag of NN.[10] Thus, we adjusted the threshold to 0.02% (102 minimum occurrences), thereby obtaining 119 frequent frames, with a precision of 82.0%. For the 30% token-based overlap (the best result for English), frames merged into 81 classes, with 79.1% precision. These precision figures are on a par with English (cf. table 7).[11] Part of this might be due to the fact that NN is still a large majority (76% of the frames). Additionally, we find that, although the frame tokens make up only 5.2% of the corpus and the types make up 15.9% of the target lexicon, those types correspond to 67.2% of the target corpus tokens.

5 Summary and Outlook

Building on the use of frames for human category acquisition, we have explored the benefits of treating contexts—in this case, frames—as categories and analyzed the consequences. This allowed us to examine a way to evaluate tagset mappings and provide feedback on distributional tagset design. From there, we explored using lexical information to combine contexts in a way which generally preserves the intended category.

We evaluated this on English and German, but, to fully verify our findings, a high priority is to perform similar experiments on more corpora, employing different tagsets, for different languages. Additionally, we need to expand the definition of a context to more accurately categorize contexts, while at the same time not lowering recall.

Acknowledgements

We wish to thank the Indiana University Computational Linguistics discussion group for feedback, as well as the three anonymous reviewers.

A Some Penn Treebank POS tags

DT	Determiner
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	*Wh*-determiner
WP$	Possessive *wh*-pronoun

[10] We use no tagset mappings for our TIGER experiments.

[11] Interestingly, thresholds of 20% and 10% result in similarly high precision.

References

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith (2002). The TIGER Treebank. In *Proceedings of TLT-02*. Sozopol, Bulgaria.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, T. J. Watson, Vincent J. Della Pietra and Jenifer C. Lai (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics* 18(4), 467–479.

Christiansen, Morten H. and Padraic Monaghan (2006). Discovering verbs through multiple-cue integration. In *Action Meets Word: How Children Learn Verbs*, Oxford: OUP.

Clark, Alexander (2000). Inducing Syntactic Categories by Context Distribution Clustering. In *Proceedings of CoNLL-00*. Lisbon, Portugal.

Clark, Alexander (2003). Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of EACL-03*. Budapest.

Dickinson, Markus (2008). Representations for category disambiguation. In *Proceedings of Coling 2008*. Manchester.

Dickinson, Markus and Charles Jochim (2008). A Simple Method for Tagset Comparison. In *Proceedings of LREC 2008*. Marrakech, Morocco.

Dubey, Amit (2004). Statistical Parsing for German: Modeling syntactic properties and annotation differences. Ph.D. thesis, Saarland University, Germany.

Elworthy, David (1995). Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*. Dublin.

Goldberg, Yoav, Meni Adler and Michael Elhadad (2008). EM Can Find Pretty Good HMM POS-Taggers (When Given a Good Start). In *Proceedings of ACL-08*. Columbus, OH.

Goldwater, Sharon and Tom Griffiths (2007). A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL-07*. Prague.

Headden III, William P., David McClosky and Eugene Charniak (2008). Evaluating Unsupervised Part-of-Speech Tagging for Grammar Induction. In *Proceedings of Coling 2008*. Manchester.

Hepple, Mark and Josef van Genabith (2000). Experiments in Structure-Preserving Grammar Compaction. In *1st Meeting on Speech Technology Transfer*. Seville, Spain.

Koo, Terry, Xavier Carreras and Michael Collins (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of ACL-08*. Columbus, OH.

Korhonen, Anna, Yuval Krymolowski and Zvika Marx (2003). Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of ACL-03*. Sapporo.

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

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*. CUP.

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

Miller, Scott, Jethran Guinness and Alex Zamanian (2004). Name Tagging with Word Clusters and Discriminative Training. In *Proceedings of HLT-NAACL 2004*. Boston, MA.

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.

Parisien, Christopher, Afsaneh Fazly and Suzanne Stevenson (2008). An Incremental Bayesian Model for Learning Syntactic Categories. In *Proceedings of CoNLL-08*. Manchester.

Redington, Martin, Nick Chater and Steven Finch (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science* 22(4), 425–469.

Schütze, Hinrich (1995). Distributional Part-of-Speech Tagging. In *Proceedings of EACL-95*. Dublin, Ireland.

Smith, Noah A. and Jason Eisner (2005). Contrastive Estimation: Training Log-Linear Models on Unlabeled Data. In *Proceedings of ACL'05*. Ann Arbor, MI.

Toutanova, Kristina and Mark Johnson (2008). A Bayesian LDA-based Model for Semi-Supervised Part-of-speech Tagging. In *Proceedings of NIPS 2008*. Vancouver.

Wang, Hao and Toben H. Mintz (2007). A Dynamic Learning Model for Categorizing Words Using Frames. In *Proceedings of BUCLD 32*. pp. 525–536.