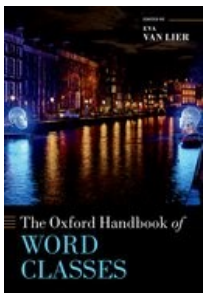

The Oxford Handbook of Word Classes

Eva van Lier (ed.)

https://doi.org/10.1093/oxfordhb/9780198852889.001.0001

Published: 07 December 2023 **Online ISBN:** 9780191887185 **Print ISBN:** 9780198852889

CHAPTER

38 Word Classes in Corpus Linguistics

Natalia Levshina

https://doi.org/10.1093/oxfordhb/9780198852889.013.34 Pages 833–850

Published: 18 December 2023

Abstract

Word classes play a central role in corpus linguistics under the name of parts of speech (POS). Many popular corpora are provided with POS tags. This chapter gives examples of popular tagsets and discusses the methods of automatic tagging. It also considers bottom-up approaches to POS induction, which are particularly important for the 'poverty of stimulus' debate in language acquisition research. The choice of optimal POS tagging involves many difficult decisions, which are related to the level of granularity, redundancy at different levels of corpus annotation, cross-linguistic applicability, language-specific descriptive adequacy, and dealing with fuzzy boundaries between POS. The chapter also discusses the problem of flexible word classes and demonstrates how corpus data with POS tags and syntactic dependencies can be used to quantify the level of flexibility in a language.

Keywords: corpora, POS tags, corpus annotation, ambiguity, distributional approach, flexible word classes, Universal Dependencies

Subject: Language Evolution, Linguistics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

38.1 Introduction

It is necessary to begin with a terminological note. The term 'word classes' is not widely used in corpus linguistics. Nouns, verbs, adjectives, and other categories are usually referred to as parts of speech (POS). The goal of POS annotation in corpus linguistics is to classify tokens (or wordforms) in context. The term parts of speech reflects this goal better than the term word classes because corpora represent *speech* as *parole*, where each unit plays a specific role determined by its context. For example, the word *play*, which can be regarded as a member of the NV word class, similar to words *love*, *walk*, or *change* (Hockett 1958: 225–227), is either a noun or a verb in a particular sentence (but see section 38.4 on how POS tags and syntactic dependencies can be used for studying the flexibility of word classes). Since most of the existing corpora represent written language rather than speech in the narrow sense, it might be even more precise to call POS 'parts of utterance', or POU.[1]

The other component of the term 'parts of speech', namely, the word 'parts', implies that we apply the tags to some speech segments. In most cases, these segments are tokens, or sequences of characters between spaces. They can be wordforms, symbols, digits, or punctuation marks. POS tags are usually assigned to these units. However, one can also assign POS tags below and above the level of tokens. For example, one can annotate a multiword expression *high school teacher* with one POS tag (noun). Alternatively, one can annotate the German orthographic word *im* as a combination of the preposition *in* and definite article *dem.* One could also use POS tags to annotate lemmas in a sentence-word in a polysynthetic language (e.g. Arkhangelskiy & Lander 2016). This is another reason why the term part of speech is better suited for corpus investigations than word classes.

The next section provides an overview of current practices in POS tagging and bottom-up induction of POS. There are many difficult decisions to make when choosing a POS tagset: the level of granularity, redundancy
p. 834 given the other layers of annotation, and so on. ↳ These criteria are discussed in section 38.3. It also demonstrates how corpora can be used to model flexibility of word classes with the help of POS tags and syntactic dependencies. Section 38.4 provides the conclusions and an outlook.

38.2 Current practices in POS tagging

38.2.1 An overview of popular POS tagsets

This section describes several popular POS tagsets used by corpus linguists. The first tagsets were created for English, because English was the first language of contemporary corpus linguistics (Atwell 2008). The first POS-tagged corpora include the Brown corpus, the Lancaster–Oslo–Bergen (LOB) corpus, the Spoken English Corpus, the Polytechnic of Wales corpus (PoW), the University of Pennsylvania corpus (UPenn), the International Corpus of English (ICE) and the British National Corpus (BNC). The tagsets used in these corpora are not limited to traditional word classes. Since the applications of corpora were not known in advance, it was tempting to add as much linguistic information as possible (Atwell 2008). A good example of such a detailed tagset is CLAWS7[2] used in well-known English corpora developed at Brigham Young University:[3] the Corpus of Contemporary American English (COCA), COHA, iWeb and some others. It includes 137 tags, plus additional tags for elements of multiword expressions (e.g. *as soon as*, *in terms of*). They combine several types of information: morphological categories (e.g. separate tags for singular and plural nouns, third person present tense verbs), individual lexemes (e.g. particle *to*, preposition *of*, existential *there*), and fine-grained subclasses of traditional parts of speech (e.g. auxiliary verbs, comparative adjectives, and degree adverbs). CLAWS7 also includes lexicogrammatical classes based on distributional and semantic properties. For example, nouns have special tags if they represent measure units (*inch*, *feet*), location (*Street*), time (*day*, *year*) or direction (*north*). These nouns occur in various idiosyncratic constructions (such as adverbial uses with zero case marking, as in *go west*, *last year*). Also, separate labels are used for letters of the alphabet, foreign words, and formulas.

An example from COCA (Davies 2008–) is shown in (1). The first column represents the wordform. The second is the lemma, and the third one is the detailed POS tag. In the CLAWS7 tagset, *mc1* stands for a singular cardinal number, *nnt1* for a singular temporal noun, *y* for a punctuation mark, *pphs1* for the 3rd person singular subjective personal pronoun, *vvd* is the past tense form of a lexical verb, *appge* is the possessive prenominal pronoun, and *nn1* is a singular common noun.

(1)

One	one	mc1
night	night	nnt1
,	,	y
he	he	pphs1
attacked	attack	vvd
his	his	appge
wife	wife	nn1
.	.	y

p. 835

After English, POS tagging has been performed for corpora in other languages, such as STTS (Stuttgart–Tübingen Tagset) for German with 54 tags[4] or the Estonian Treebank POS tagset with an impressive number of 579 tags.[5] There have been attempts to standardize corpus annotation in different languages. One should mention here an important initiative called EAGLES (Expert Advisory Group on Language Engineering Standards) supported by the European Union with the aim of developing comparable and interchangeable technologies for the multilingual European community. The experts' recommendation was to use an obligatory set of classes including noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unique/unassigned (for small and very specific categories, such as negative particles), residual (e.g. foreign words or mathematical formulae) and punctuation (Leech & Wilson 1999).

These categories are similar to tagsets used for multilingual corpora nowadays. For example, the Universal POS tags used in the Universal Dependencies corpora (Zeman et al. 2020), include six tags for open-class words: ADJ (adjective), ADV (adverb), INTJ (interjection), NOUN (common noun), PROPN (proper noun) and VERB (lexical verb). There are also eight tags for closed-class words: ADP (adpositions), AUX (auxiliaries and copulas), CCONJ (coordinate conjunctions), DET (determiners), NUM (numerals), PART (particles), PRON (pronouns) and SCONJ (subordinate conjunctions). Three remaining tags include PUNCT (punctuation), SYM (symbol) and X (other, including foreign words, URLs, etc.).[6] Morphological information is language-specific and provided in a separate layer.

Similarly, the Google tagset for cross-linguistic data includes 12 tags (Petrov et al. 2012): nouns, verbs, adjectives, adverbs, pronouns, determiners (including articles), adpositions, numerals, conjunctions, particles, punctuation marks and the category 'X', which is a catch-all for other categories such as abbreviations or foreign words. This tagset has been used for annotation of the Google Books Ngram Corpus in several major languages.[7]

Interestingly, these tagsets, which have been developed for NLP applications in diverse languages, are closer to the traditional parts of speech derived from traditional Latin grammatical categories (noun, verb, adjective, preposition, pronoun, adverb, conjunction, and interjection) than the first tagsets in corpus linguistics.

38.2.2 Methods of top-down automatic tagging and evaluation

One can annotate data by hand, but for large corpora this is not feasible. There exist diverse algorithms that provide POS tags automatically.

p. 836 The main problem of POS tagging is disambiguation, or determining the POS of homographs like *play* (verb or noun) in the context. In the beginning of POS tagging in the 1950s and 1960s, one had to manually create rules for this purpose. For example, a token could not be a verb if it was immediately preceded by an article. According to Voutilainen (1999: 11), the first system was created at the University of Pennsylvania in the late 1950s. The tagging part contained a small disambiguator with 14 ordered context rules.

Nowadays, it is more common to use data-driven approaches, where one usually takes a manually annotated training corpus and creates a language model, which takes into account diverse statistical associations in the corpus. The tagger then uses this statistical model to select the tags with the highest probability in a given context. Already early implementations reached spectacular accuracy. For example, the first CLAWS parser (Constituent-Likelihood Automatic Word Tagging System, version 1) developed for the LOB corpus back in the 1980s had accuracy above 95% (Voutilainen 1999). Of course, the procedure is rarely fully automatic. One relies on lexicons, idioms lists, special rules, and other resources. Often, human annotators check manually the tokens that remain ambiguous.

Different surface cues can be useful for this purpose: *n*-grams (sequences of *n* words) preceding and following the target word (e.g. Brown et al. 1992); orthographic criteria, such as capital letters, digits, internal hyphens, apostrophes (e.g. Yatbaz et al. 2012); and morphological information (e.g. Clark 2003), where the algorithm tries to find suffixes or prefixes in words and combine the words that share common morphology. More sophisticated approaches like neural networks and hidden Markov models are able to take into account long-distance patterns. A combination of different criteria usually helps to improve the performance (Christodoulopoulos et al. 2010; Yatbaz et al. 2012). There are also successful approaches that are based on cluster prototypes (e.g. Haghighi & Klein 2006).

The performance of a tagger is usually evaluated on a test corpus with the help of several popular measures: correctness, ambiguity, precision and recall (van Halteren 1999). Correctness is simply the proportion of correct tags appropriate in the context relative to all tokens. Ambiguity is the average number of tags per token (see more in section 38.3.4). Precision shows how many tokens with a given tag (e.g. 'NOUN') are tagged correctly. Finally, recall is used to measure how many tokens that should have a certain tag are indeed tagged so. High performance numbers, however, do not necessarily mean that the tagging is useful because the usefulness also depends on the quality and quantity of information conveyed by the tags. One should also be aware that the tagger's performance may differ from one text type to another.

Many current taggers have performance with accuracy above 97% per token, which is at least as good as human interrater agreement (Manning 2011). This means that above 97% tokens are analysed correctly. However, if one computes the number of correctly analysed sentences, the same taggers have sentence accuracies around 55%–57%, a rather modest score, which has to do with the fact that some units are extremely frequent and ambiguous, e.g. *that*, which can function as a pronoun, determiner, complementizer, or even adverb, as in *She is not that crazy* (Manning 2011). In some cases, there may not be good conventions in linguistic descriptions. For example, the word *worth* has properties of both an adjective and a preposition governing a nominal phrase, as in *It's not worth the effort*. See section 38.3.4 on fuzzy boundaries between POS. Here, the prospects of POS tagging depend very much on improved descriptive linguistics (Manning 2011).

38.2.3 Bottom-up induction of POS

The approaches discussed above presuppose some ‘gold standard’, which is used to evaluate the accuracy of the labels. A different approach is induction of POS categories from the data in a bottom-up way. One such attempt is Wälchli (2008), who uses unlemmatized translations of the New Testament to perform unsupervised clustering based on Biemann’s (2006) graph-theoretic method of Chinese whispers. Wälchli reports interesting cross-linguistic differences. Some languages yield few distributional classes, which emerge as a result of clustering (e.g. French has only 14 clusters), while others have hundreds (e.g. Finnish with 108 clusters). Some languages yield more formal classes, like French (e.g. finite verbs 3SG, auxiliaries, feminine nouns singular), whereas Vietnamese has 100 clusters with many semantic groups, such as body parts, animals and, surprisingly, even words related to doors and walls. It is pointed out that the distributional analysis based on the neighbouring words alone does not suffice for languages with rich morphology. For example, in synthetic languages some clusters represent case forms of nouns and pronouns or personal forms of verbs (e.g. Finnish, Hungarian, Latin, and Turkish). In particular, the accusative singular form *manum* ‘hand’ in Latin ends up in one cluster, and the ablative singular form *manu* in another. One needs some morphological information in order to improve the clustering.

The bottom-up approach to POS induction has been particularly important for theory of child language acquisition. More exactly, researchers usually use the term ‘syntactic categories’. A fundamental question is how children perform induction of adult-like categories from the ambient language. Child language acquisition research was influenced strongly by Chomsky’s (1972) idea known as the ‘poverty of stimulus’ argument: the data that children are exposed to is insufficient for learning grammar only from ambient language, therefore a large part of a child’s grammar is innate.

However, there have been quite a few corpus-based studies showing that syntactic categories can be inferred from linguistic input data (see Diessel 2009). For example, Redington, Chater, & Finch (1998) use the CHILDES corpus (MacWhinney & Snow 1985) to perform hierarchical clustering of words based on the words that occur on the left and on the right. They obtained such interpretable classes, as proper nouns, adjectives, common nouns, prepositions, one large cluster of verbs, a cluster of determiners and possessive pronouns. They found that two words on the left and on the right together provide the clusters that are the most similar to the benchmark (12 classes from the Collins Cobuild Lexical Database). Similarly, Mintz, Newport, & Bever (2002) showed that co-occurrence patterns with surrounding words allow one to successfully categorize the majority of nouns and verbs. All this means that distributional information is a powerful cue for learning syntactic categories. In addition, Moran et al. (2018), who compare syntactic frames (frequently occurring nonadjacent sequences of words, e.g. *I* X *books*, as in *I read books*) and morphological frames (sequences of morphemes, e.g. *is* X*ing*, as in *is sleeping*), find that only the latter can serve as highly accurate cues cross-linguistically for learning of word classes. Therefore, morphological distributional information seems to be at least as relevant as the information about the surrounding words. This means that the term ‘syntactic categories’ is not very fortunate, and should be regarded as a manifestation of Anglocentrism in theoretical linguistics.

p. 838

38.3 Part-of-speech annotation: Conflicting requirements and difficult Decisions

38.3.1 What makes a good part-of-speech annotation?

The goal of POS annotation is to maximize the usability of a corpus for its target users, who belong to different communities with diverse interests: from lexicographers and teachers to NLP researchers and software developers. Their needs should be reflected in the choice of POS tagsets. For example, the LOB corpus has a fine-grained, complex tagset because it was developed primarily for the purposes of English language teaching and research, while the UPenn (University of Pennsylvania) corpus has a smaller tagset, which is more convenient for Machine Learning applications (Atwell 2008).

The criteria of useful POS annotation include the following, which are often in conflict.

1. The level of granularity should allow for extraction of linguistic patterns of interest with maximal precision. As a result, annotation should be fine-grained. At the same time, the search for traditional broad categories, such as a noun or a verb, should be possible without too much effort. Moreover, the tags should be easy to remember. Therefore, the list of tags should be relatively short.

2. The distribution of linguistic information across different layers should be efficient. This means that POS annotation should not overlap with the other layers, such as morphological features and syntactic roles. At the same time, it should provide sufficient information for each layer to be used independently.

3. The annotation should strive for the ideal 'one token—one POS tag'. At the same time, it should be sufficiently flexible in order to take genuinely ambiguous uses into account.

4. POS annotation should reflect the existing standards and traditions in language description, if there are any, to make it easier for users to search for different categories. And yet, it should, ideally, be transferrable to other languages and varieties.

As one can already see from these contradictory desiderata, development of a perfect POS tagset is similar to attempts of squaring a circle. The remaining part of this section provides illustrations of how several well-known corpora deal with these challenges.

38.3.2 The level of granularity

As was discussed in section 38.2.1, there is substantial variation in the size of existing tagsets. The choice of tags depends on multiple factors, such as the application, the state of tagging algorithms and the performance expected by the tagger (Paroubek 2007). In particular, developers may simplify tags because they cannot achieve a required level of accuracy with the tools they have at disposal. For instance, the performance of a POS tagger for French could be only improved by excluding the gender information from the tags of nouns and adjectives (Chanod & Tapanainen 1995).

p. 839

The large and detailed tagsets like CLAW7 are fine-tuned to the grammatical peculiarities of a language and help to search for specific words and constructions with high precision. For example, if one wants to find examples of *What a(n) NOUN!* (e.g. *What an idiot!*), one can use the wordform *What* followed by the indefinite article *at1* and a singular common noun *nn1*.

In most tagsets with detailed information, the first positions in the tags (e.g. *n* in *nn1* or *v* in *vvd*) represent the traditional parts of speech (i.e. nouns or verbs). In CLAWS7, each major part of speech begins with one unique letter. For example, we can find all adjectives by searching for a tag that starts with the letter 'J'.

Some tagsets are less convenient in that regard. For example, in the BNC the C5 tags of all nouns start with 'N'. However, if one wants to find adjectives, it is necessary to specify two letters, 'AJ'.[8] If one simply looks for 'A**', they will also get articles ('AT0') and adverbs ('AV*'). At the same time, the XML version of the BNC provides an additional layer of annotation, which corresponds to the traditional parts of speech. See an example in (2). Each word is on a separate line. The tags inside of the annotation XML tags <w... > before the wordform represent the coarse-grained POS tag 'pos' (e.g. pronoun, verb, substantive), the lemma 'hw', and the fine-grained POS tag from the 'c5' tagset, e.g. *PNP*—personal pronouns, *VVZ*—3rd person singular present tense of a verb, *NN1*—common singular noun, etc.

(2) Sentence: *She gazes at herself in wonder.* (FBo)
Annotation:
<s n='49'>
<w pos='PRON' hw='she' c5='PNP'>She </w>
<w pos='VERB' hw='gaze' c5='VVZ'>gazes </w>
<w pos='PREP' hw='at' c5='PRP'>at </w>
<w pos='PRON' hw='herself' c5='PNX'>herself </w>
<w pos='PREP' hw='in' c5='PRP'>in </w>
<w pos='SUBST' hw='wonder' c5='NN1'>wonder</w>
<c c5="PUN">.</c>
</s>

A more parsimonious way is to completely separate the fine-grained features such as number and person, from the major classes. This approach is implemented in the Universal Dependencies corpora (Zeman et al. 2020). An example from the English EWT corpus is provided in (3).[9] Each token is represented by several columns: token ID, token, lemma, Universal Part of Speech (UPOS), a finer-grained POS tag, morphological features, such as mood, tense, number, case, person, and degree of comparison, as well as the ID of the syntactic head, and finally the syntactic dependency relation (root, object, subordinator, etc.). Importantly, the UPOS tags are separated from the morphological features.

p. 840

(3) # sent_id = email-enronsent31_01-0012
text = Take care and hope to hear from you soon.

1	Take	take	VERB	VB	Mood=Imp|VerbForm=Fin	0	root
2	care	care	NOUN	NN	Number=Sing	1	obj
3	and	and	CCONJ	CC	_	4	cc
4	hope	hope	VERB	VBP	Mood=Ind|Tense=Pres|VerbForm=Fin	1	conj
5	to	to	PART	TO	_	6	mark
6	hear	hear	VERB	VB	VerbForm=Inf	4	xcomp
7	from	from	ADP	IN	_	8	case
8	you	you	PRON	PRP	Case=Acc|Person=2|PronType=Prs	6	obl
9	soon	soon	ADV	RB	Degree=Pos	6	advmod
10	.	.	PUNCT	.	_	1	punct

As mentioned, cross-linguistic tagsets, such as UPOS and Google tagsets, are coarse-grained. This is understandable, since the fine-grained categories are not (fully) applicable across different languages. It is not excluded that future development of tagsets beyond well-described European languages will result in even smaller and more general tagsets.

38.3.3 Distribution of information across different layers of annotation

As mentioned above, grammatical information can be distributed over several layers. This solves the problem of granularity. The user can combine information from different layers in any way to get more or less detailed descriptions. There is another problem here, however. Ideally, these layers should provide unique information, i.e. they should be orthogonal.

According to the famous account by Croft (2001: ch. 2), the core word classes (nouns, verbs, and adjectives) have two sides: semantics (e.g. objects, properties, and actions) and 'information packaging', such as reference, modification, and predication, which corresponds to syntactic constructions where they can be used. At the same time, because of their semantic component, they are not reduced to their syntactic roles.[10]

Croft et al. (2017) argue that semantics and information packaging should be disassociated in corpus annotation. They suggest to perform a simplification of the syntactic dependencies used in the UD corpora. The dependencies represent syntactic functions, such as subject, object, predicate, complement clause, etc.[11] One could group, for example, determiners, numeric modifiers, and adjectival modifiers of nouns under a more general tag 'modifier'. The differences between them can be captured by the Universal POS (UPOS) tags, which were discussed in sections 38.2.1 and 38.3.2. In that case, the annotation scheme would be more parsimonious.

p. 841 However, this has not been implemented. The reason is practical. Some users (e.g. syntacticians) prefer using the dependencies, while some others use only the POS tags (which are perhaps more widely used in corpus linguistics). This annotation is therefore redundant, but practically useful.

But how serious is the problem of redundancy in actual corpus data? As an illustration, let us have a look at two layers in the Universal Dependencies corpora: UPOS and syntactic dependencies. According to Croft et al., these two types of information should not overlap, for the annotation schema to be maximally informative. One possible way of measuring this systematically is to compute Shannon's (1948) entropy for each UPOS tag in a corpus across different syntactic dependencies. If the entropy is high, we will not be able to predict the syntactic role from the part of speech. This means that each of the layers carries unique information. In contrast, if the entropy is low, this means that a particular UPOS tag is strongly associated with one or two syntactic functions, and the annotation has high redundancy.

Table 38.1 provides an illustration. It displays a subset of the frequencies of tokens annotated as NOUN in different syntactic roles in the corpus of Afrikaans in the Universal Dependencies corpora (version 2.5). These frequencies are divided by the total number of all tokens with a particular tag (here: NOUN) to obtain the probabilities.

Table 38.1 Some frequencies of common nouns as syntactic dependencies in the Afrikaans Universal Dependencies training corpus (v 2.5)

Universal POS	nsubj	obj	iobj	csubj	ccomp	xcomp	obl	vocative	expl
NOUN	903	1,573	42	0	0	1	1,557	0	0

The results averaged for each POS in the UD corpora are displayed in Figure 38.1. It shows that common nouns (NOUN) have the highest entropy on average. This means that they are the most diverse syntactically, and it is the most difficult to predict the syntactic role of a noun. They are followed by pronouns (PRON), lexical verbs (VERB) and proper nouns (PROPN). On the other end of the continuum are the subordinate (SCONJ) and coordinate (CCONJ) conjunctions, followed by adpositions (ADP). An examination of the syntactic dependencies reveals that these three Universal POS tags nearly always co-occur with the roles of subordination markers (*mark)*, coordinating conjunctions (*cc*), and case markers (*case*), respectively. Adverbs (ADV) are often adverbial modifiers (*advmod*), whereas interjections (INTJ) are annotated as discourse elements (*discourse*). Auxiliaries (AUX) are syntactically annotated as auxiliaries (*aux*) and copulas (*cop*). Particles (PART) are adverbial modifiers (*advmod*), subordinating markers (*mark*) and discourse elements (*discourse*).

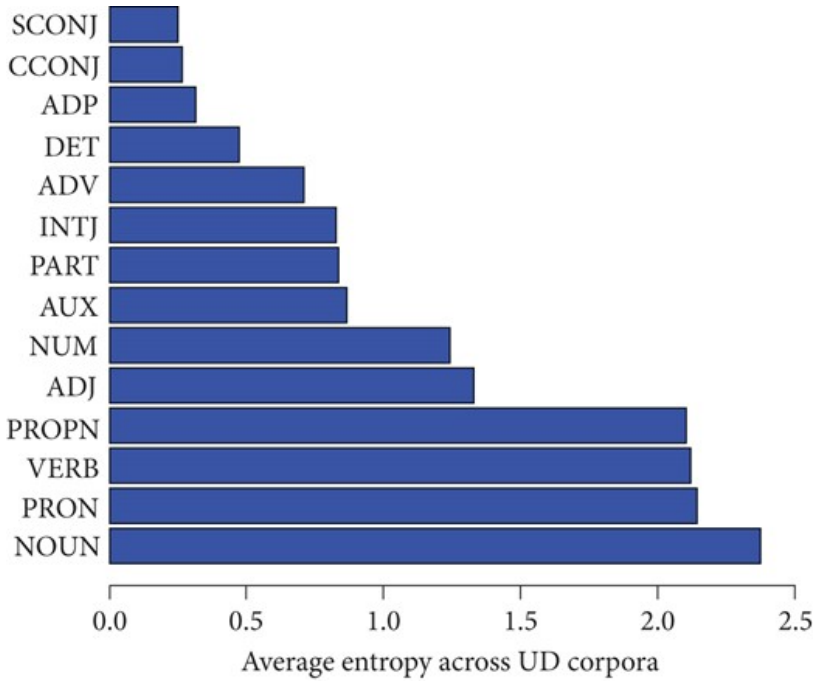

Figure 38.1 Average entropy scores of Universal POS tags regarding the syntactic dependencies in the Universal Dependencies corpora

These results demonstrate that most function words are strongly associated with particular syntactic dependencies. This means a maximal overlap between UPOS and dependencies for these word classes. This is hardly surprising. Function words are defined by their syntactic role. Nouns, verbs, and adjectives have less
p. 842 redundancy. These POS are in ↳ fact doing well in comparison to the others. So, if one wanted to make any improvements in terms of redundancy, they would need to start with function words.

38.3.4 Fuzzy boundaries between parts of speech

Ideally, every token in a corpus should have one POS tag. This rule can be violated in three cases. First, the boundaries between parts of speech in a language can be fuzzy (e.g. Croft 2001: 103). A descriptively adequate POS annotation should be sufficiently flexible to take that fuzziness into account. Second, sometimes there is simply not enough contextual information for disambiguation between two categories. For example, it may be unclear whether the word *fire* in a short exclamatory sentence *Fire!* is a noun or a verb (Cloeren 1999: 48). The first and the second cases represent genuine ambiguity. Third, there can be not enough information for the automatic tagger to make a decision, but this ambiguity can be resolved by a human annotator.

Some well-known corpora, such as the BNC and COCA, have ambiguity tags. As an illustration, consider the BNC tagset (see section 38.3.2), which contains ambiguous tags, such as VVD–VVN. This means that the automatic tagger was unable to decide whether the word is a VVD (past tense verb) or a VVN (past participle). The two possibilities are left for the users to disambiguate.

Ambiguity tags constitute approximately 4.7% of the BNC tags (excluding punctuation tags). Table 38.2 displays the top ten most frequent ambiguity tags in the XML edition of the corpus. Which type of ambiguity do they
p. 843 represent? ↳

Table 38.2 Top 10 ambiguity tags in the British National Corpus

Tag	Meaning	Frequency	Example
NN1–VVB	Singular common noun or finite base form of lexical verb	528,014	*pay* cheques or slips
AJ0–NN1	General/positive adjective or singular common noun	421,584	the *chief* executive
NP0–NN1	Proper noun or common singular noun	347,428	I'm looking for *Bill*.
PRP–AVP	Preposition or general adverb	245,875	Put it *on* the reverse.
NN1–AJ0	Singular common noun or general/ positive adjective	238,719	restrictive *patient* choice
VVB–NN1	finite base form of lexical verb or singular common noun	218,459	*Control* in this context has been defined by …
VVN–VVD	past participle or past tense form of lexical verb	206,468	A survey *conducted* in the United States discovered …
VVD–VVN	past tense or past participle form of lexical verb	190,747	Completely *ignored* it!
NN1–NP0	Common singular noun or proper noun	156,747	Any *Don* Juan can say he loves you.
VVG–AJ0	The -ing form of lexical verb or general/positive adjective	118,329	all seeming big and *threatening*

A closer look at the corpus data reveals some genuine cases of ambiguity. In particular, the adjective—noun choice is often very tricky. Examples are *its opposite*, *its excess population*, *some wild-eyed back-country messiah*, and *the protestant parts*. Another case of genuine ambiguity is VVG–AJ0, which represents an *ing*-participle or an adjective, e.g. *all these working people*, *developing countries*, *preceding and following period*, *measuring instruments*, *be terrifying*. These examples seem to illustrate the fuzzy boundaries between the semantic properties of POS. It is understandable that these words have ambiguity tags.

However, most of the examples in the corpus are simply due to problems with the automatic annotation. For example, the noun *mind* in *bearing in mind* is ambiguous with a verb (NN1–VVB) in one context. For a human annotator, these decisions are straightforward. Ideally, each of those tags should be manually checked and corrected, so that only the genuine cases remain. Unfortunately, the difference between genuine ambiguity and the algorithm's deficiency is rarely reflected in available corpora.[12]

One should also mention here a very different case, when the ambiguity of POS tags is less obvious. For example, in the EAGLES tags for Romance languages (see section 38.2.1), there is a value *C* for gender meaning
p. 844 'either F or M'. Nouns like Spanish *estudiante* 'a male or ↳ female student' are then annotated as having no gender, although in fact they do have a specific gender in a given context. The problem is that it is very difficult for a parser to annotate the gender correctly, which worsens the parser's performance.[13] Therefore, when evaluating POS ambiguity, one should also consider the level of detail provided by the tagset.

38.3.5 Cross-linguistic usability and traditional language-specific descriptions

Recently there has been strong interest in developing and applying uniform annotation schemas for different languages. The practical benefits of a standardized tagset are obvious: one can interchange and reuse resources and develop corpora and tools that can be used globally (Leech & Wilson 1999). Creating a tagset applicable to many diverse languages is a challenging task. The creators should strike a balance between the usefulness of the tagset for a maximum number of languages and its descriptive adequacy for any specific language. Let us consider these criteria in greater detail, using the Universal Dependencies corpora as an illustration.

The first criterion, the usefulness of the tagset for corpora in different languages, can be evaluated by counting the number of tags that are used in all corpora and those that are not used. Out of 150 corpora available at the moment of writing, 46 corpora contain all seventeen POS tags. 52 corpora contain all but one, and 30 corpora contain all but two, which means that more than 85% of the corpora have all or almost all tags. The total counts are shown in Table 38.3. Note that the missing tags are usually technical (see below).

Table 38.3 Number of unused tags (maximum: 17)

Number of unused tags	0	1	2	3	4	5	6	9
Number of corpora	46	52	30	13	5	5	1	1

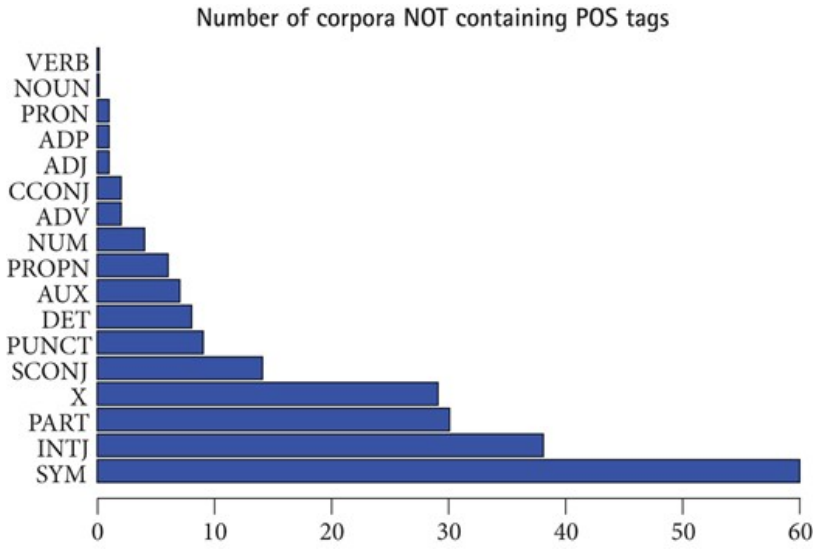

Figure 38.2 Number of corpora in the Universal Dependencies collection that do not have individual universal part-of-speech tags

Let us now look at the distribution of each individual Universal POS tag. Figure 38.2 displays the number of corpora in which a particular tag was not used. The tags VERB and NOUN are used in all corpora. They are followed by PRON (pronouns), ADP (adpositions), and ADJ (adjectives). The tags SYM (symbol), INTJ (interjections) and X (foreign words and other dubious cases) are the least commonly used, probably because they are on the periphery of a linguistic system, and do not receive much attention. Another reason may be that some corpora do not contain tokens representing these classes (e.g. one will not find many interjections in news reports, or symbols in spoken corpora). Among the traditional parts of speech, the ones that are used the least frequently are PART (particle), SCONJ (subordinate conjunctions), and DET (determiners). Some corpora do not annotate proper nouns (PROPN), AUX (auxiliaries), and PUNCT (punctuation).

p. 845 The corpus with the largest number of unused tags is Warlpiri UFAL. It has nine categories that are not used. This may be at least partly due to data sparseness because the corpus is very small and contains grammar examples. The Tagalog corpus, which has five tags missing, is similar in that respect. For example, CCONJ (coordinate conjunctions) and NUM (numerals) are missing in the Warlpiri and Tagalog corpora, but these word classes are also missing in the English translations, which are provided in the files. This seems to be an artefact of the source data and is likely to change when more natural texts are added.

Some of the tags may be left out for a good reason. For example, the corpus of Swedish Sign Language does not contain tokens annotated with the punctuation tag PUNCT. ADP (adpositions) are missing in one of the Korean treebanks (the parallel treebank PUD), but the status of case and topic markers (particles), which can be interpreted as adpositions, is not straightforward in Korean. There are no ADJ (adjectives) in Classical Chinese (the treebank Kyoto). The functionally similar words are annotated as VERB. According to the corpus creators, 'adjective usages of verbs were not specialised as adjectives at that era' (https://universaldependencies.org/lzh/index.html).

Some missing tags are difficult to explain, however. For example, the Latin-Perseus corpus lacks PROPN (proper nouns), which are annotated as common nouns instead. PRON (pronouns) are absent in the Vietnamese VTB corpus. For some reason, they are coded as proper names. SCONJ (subordinate conjunctions) are missing in two Portuguese corpora, where they appear as CCONJ (coordinate conjunction) and in one Indonesian corpus, where they are annotated as adpositions. There is obviously room for improvement in such cases.

Overall, if we compare Indo-European with the other languages represented in the UD corpora, we will see that there are more non-Indo-European corpora with missing tags, as far as most traditional parts of speech are
p. 846 concerned. The frequencies of missing tags in ↳ Indo-European and all other languages are shown in Figure 38.3. This difference becomes even more striking if we consider the fact that the total number of non-Indo-European corpora is almost twice as small as the number of Indo-European ones (only 57 vs 97). This means that the Universal POS annotation does have an Indo-European bias. There is a caveat, however. Many non-Indo-European corpora are very small and non-naturalistic, so some of these biases may disappear in the future, as more diverse texts become available.

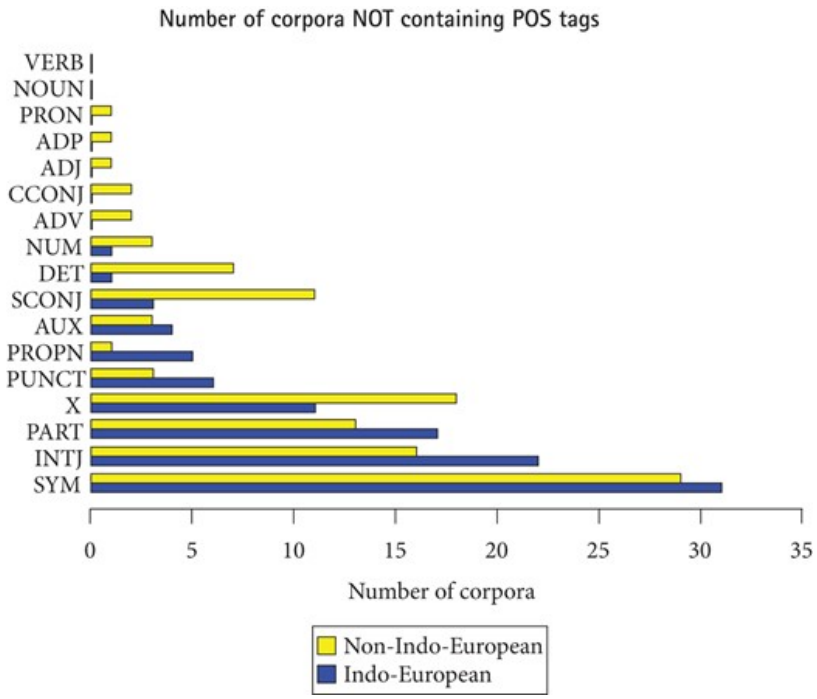

Figure 38.3 Number of corpora in the Universal Dependencies collection without individual part-of-speech tags: Indo-European vs Non-Indo-European languages

The second criterion of POS tags usability is close correspondence between POS annotation and traditional language-specific descriptions. It is easier to search for constructions in a specific language when the tags are familiar from previous language descriptions. The Universal POS tags often deviate from this ideal. For example, cardinal numerals are annotated as numerals, but ordinal numerals are treated as adjectives because ordinal numerals are morphologically and syntactically similar to adjectives.

Another example is the annotation of particles (PART). For instance, Japanese has many function words that are traditionally called particles, but not all of them have the PART tag in the Universal Dependencies corpora of Japanese. Some of them do, e.g. the question particle *ka.* Others (e.g. *ni* and *no*) are parallel to adpositions in other languages and are tagged as ADP. Similarly, the Russian conditional/subjunctive particle *by* is coded as an auxiliary AUX because it accompanies the lexical verb of a verb phrase and expresses grammatical distinctions
p. 847 not ↳ carried by the lexical verb. This is very unconventional, since *by* is no longer perceived as a form of the verb *byt'* 'be' and has lost inflection and phonological substance. However, this choice makes sense from the cross-linguistic perspective.

In some cases, the cross-linguistic comparability is not achieved, as in the case of the boundary between determiners DET and pronouns PRON, which receives a lot of attention in the annotation guidelines.[14] In case of possessive pronouns, the developers advise annotation in accordance with language-specific properties of this subcategory. If possessive pronouns are more likely to be used attributively (modifying a noun phrase) than substantively (replacing a noun phrase), and if their inflection is similar to that of adjectives and distinct from nouns, they are annotated as DET. If they are more like normal personal pronouns (e.g. in the genitive case), they should be tagged PRON. One may wonder if this approach is optimal. It might be useful to make a clearer distinction between the Universal POS tags, which should be maximally universal, and language-specific tags, which are often available as a separate layer of annotation (i.e. as XPOS tags in the UD corpora).

Another question is how these guidelines are implemented. In most cases, the decisions are consistent with the rules. For example, possessive pronouns in the corpus English-EWT are annotated as PRON. This is reasonable because these words lack agreement with head nouns. In Czech-PDT and Russian-SynTagRus, possessive

pronouns are annotated as DET. Again, this choice is justified because they are similar to adjectives in declension and agreement. At the same time, some annotation choices are difficult to explain, especially if different decisions are made for corpora representing the same language. For example, possessive pronouns are annotated as PRON in the Latin-Perseus corpus, but as ADJ in Latin-ITTB and PROIEL.

38.4 POS and flexible word classes

The boundaries between content words (i.e. nouns, adjectives, verbs, and adverbs) vary cross-linguistically with regard to their morphosyntactic behaviour (Hengeveld 1992b; van Lier & Rijkhoff 2013; see Chapter 6 in this volume). For instance, Croft (2001: 69) points out that English exhibits substantial flexibility. Many lexemes, especially those belonging to the basic vocabulary, can be used in different functions (e.g. *a big house*, *a big* and *think big*). Another example is Tukang Besi, an Austronesian language spoken in Indonesia. In that language, words that can be used as adnominal modifiers can also be used freely, without additional coding, as arguments (the prototypical function of nouns) and as predicates (the main function of verbs) (Chapter 18 in this volume).

As was mentioned in the beginning of this chapter, POS are used to classify tokens and related syntagmatic units in context, rather than words as paradigmatic types. Therefore, POS, unlike word classes, cannot be flexible. In flexible languages, POS will be determined based on the syntactic roles. Each of the instances of big
p. 848 in *a big house*, *a big* and *think big* has ↳ a different POS tag (namely, adjective, noun, and adverb, using a coarse-grained annotation schema).

At the same time, POS tags can be easily aggregated across individual wordforms like *big*, so that one can obtain paradigmatic word classes for the purposes of studying word-class flexibility. The advantage of using corpora is that they can help us evaluate how systematic word-class flexibility is in a language. In other words, can the formal overlap be observed across the whole lexicon, or only in a few selected items? Below I show how this can be done for the adjective—adverb distinction based on the Universal Dependencies corpora of six languages.

The data were taken from the Universal Dependencies corpora of Chinese, Dutch, English, French, German, and Indonesian.[15] All tokens (wordforms) with POS tags ADJ (adjective) or ADV (adverb) were collected with the help of a Python script, and the frequencies of these tokens in the function of *amod* (adjectival modifiers of nouns, e.g. *hard work*) or *advmod* (adverbial modifiers of verbs or adjectives, e.g. *work hard*) were collected.

For each wordform, the log-odds ratio of being attracted towards the *amod* or *advmod* role was computed based on the following formula:

$$(4) \qquad \log OR = \log\left(\left(amod_{token} / advmod_{token}\right) / \left(amod_{other} / advmod_{other}\right)\right) =$$

$$= \log \frac{amod_{token} * advmod_{other}}{advmod_{token} * amod_{other}}$$

where $amod_{token}$ is the frequency of a token in the *amod* position, $advmod_{token}$ is the frequency of the same token in the *advmod* function, $amod_{other}$ is the frequency of occurrences of all other tokens in the *amod* position, and $advmod_{other}$ is the frequency of occurrences of all other tokens in the *advmod* position. By including the frequencies of the other tokens, we control for the differences between the corpora in the relative frequencies of *amod* and *advmod*. Log-odds ratio is a traditional measure of attraction between two categorical variables (Levshina 2015: ch. 9). Here, the variables are the token identity and the syntactic role. In order to deal

with zero frequencies, which would cause division by zero, a small amount (0.05) was added to each of the frequencies. Taking the logarithm allows us to centre the value around zero, such that positive values mean attraction of the token towards the *amod* function, and negative ones towards the *advmod* function. The more extreme the values, the more specialized a form is in one or the other function.

Finally, I took absolute values of the *logOR* scores. The higher the absolute scores, the more in general the words are specialized in one or the other role, and the less flexible the language. The lower the scores, the more overlap there is between the *amod* and *advmod* tokens, and the more flexible the languages are, as represented by the corpora. Only the frequencies of the wordforms that occur ten times or more in either role were
p. 849 analysed, in order to avoid unjustifiably high scores that can arise due to data sparseness. ↳

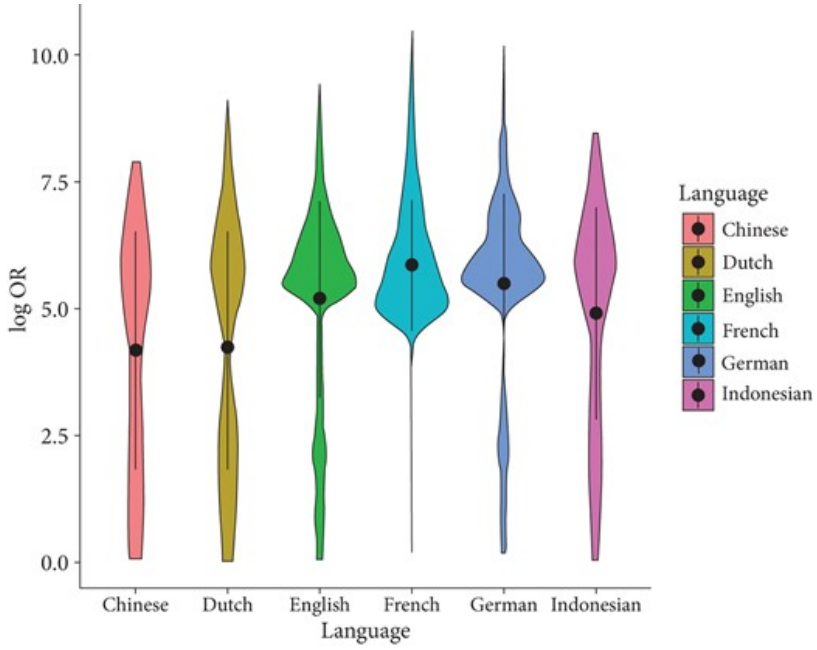

Figure 38.4 Absolute log-odds ratios of wordforms in six Universal Dependencies corpora, with the means (dots) and one standard deviation from the mean (black lines)

Figure 38.4 shows the distributions of the absolute log-odds ratios in the six corpora. The thickness of the figures represents the density (i.e. the number of tokens per interval of logOR values). The distributions are weakly bimodal. This means that the wordforms tend to be either flexible, or non-flexible. However, most wordforms seem to have high values, which means that they are attracted to one of the roles. The wordforms strongly attracted to *amod* are often adjectives representing size, novelty (old/new), nationality, as well as ordinal numerals, which are annotated in the UD as adjectives. Here also belong inflected adjectives, if they exist. As for the wordforms strongly attracted to *advmod*, they are highly frequent semi-function words *there*, *also*, *now*, *too*, or *then*. For example, in German the words with high scores are inflected adjectival forms, e.g. *gute*, *ehemaligen*, *verschiedene*, or highly frequent adverbs, such as *noch*, *auch*, *sehr*, *nur*, and *heute*.

Low scores mean that the relative frequencies of the wordforms as *amod* and *advmod* are similar to the total proportions of *amod* and *advmod* in the corpus. For example, in English, these include directions (e.g. *south*), dimensions (e.g. *wide*, *long*) and other properties (e.g. *hard*, *quick*, and *online*). In German, these are words like *nahe*, *super*, *unmittelbar*, *kurz*, *ganz*, *speziell,* which are often used as postnominally, e.g. *die Lage unmitterbar vor der Küste* 'the location directly in front of the coast'.

The distributions overlap greatly, which indicates that the flexibility is a matter of degree, although there are also subtle differences. As expected, Chinese and Dutch tend to have more wordforms with low scores and high

p. 850 flexibility. English, German, and particularly ↳ French have in fact relatively few overlapping forms. This is not surprising, because French and German adjectives agree with the noun in many contexts, and English and French have special markers for adverbs (*-ly* and *-ment*, respectively). Interestingly, Indonesian does not exhibit very high flexibility—a finding that needs further investigation. These results suggest that flexibility is a gradient phenomenon, which can be measured with the help of corpus data. POS tags combined with syntactic dependencies can provide a source for studying word classes flexibility cross-linguistically.

38.5 Conclusions

The present chapter has discussed current practices and challenges in POS annotation. Speaking generally, the approaches to word classes in linguistic literature and to POS tagging in corpus-linguistic practice are very different. First, POS tags are always assigned in context, whereas word classes are sometimes treated as paradigmatic generalizations. Second, the annotation decisions made by corpus linguists depend to a large degree on practical needs and challenges, rather than on theoretical considerations. For example, the level of detail provided in the POS tags depends on the presence of other annotation layers (e.g. syntactic dependencies and morphological features), feasibility of automatic tagging of particular features and other factors. Interestingly, while the early tagsets for English were very fine-grained so as to maximize the usefulness of POS tags for various applications, the contemporary tagsets developed for multilingual NLP are closer to the traditional word classes. This suggests that the traditional categories reflect relevant properties, after all.

Bridging this gap would be beneficial to both communities. Corpora can provide a testing ground for pertinent theoretical issues, such as word classes flexibility, at the same time forcing linguists to formulate more precise, empirically testable hypotheses. On the other hand, annotation inconsistencies, especially in multilingual corpora, suggest room for improvement. Some nudging on the part of typologists may help to improve the comparability of cross-linguistic tagsets. As the interests of corpus developers expand to the languages that differ substantially from Standard Average European, thanks to such international initiatives as the Universal Dependencies project, we will soon see how universal the current 'universal' tagsets are in reality. We are living in exciting times.

Notes

1 I thank Michael Cysouw for this idea.
2 http://ucrel.lancs.ac.uk/claws7tags.html
3 https://www.english-corpora.org/
4 https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/#id-cfcbf0a7-0
5 https://www.cl.ut.ee/korpused/morfliides/seletus/
6 See https://universaldependencies.org/u/pos/index.html.
7 https://books.google.com/ngrams
8 http://www.natcorp.ox.ac.uk/docs/c5spec.html
9 Some annotation layers, which are irrelevant for the discussion, are not shown.
10 An additional source of entropy is the fact that the reference function performed primarily by nouns and pronouns is instantiated by different syntactic arguments (subject, direct object, indirect object, and oblique).
11 See https://universaldependencies.org/u/dep/index.html.
12 Quite tellingly, a document with guidelines for EAGLES-style annotation says that dealing with genuine ambiguity is not a matter of great priority: https://home.uni-leipzig.de/burr/Verb/htm/LinkedDocuments/annotate.pdf (p. 17).
13 I thank Maarten Janssen for bringing up this important point.
14 See https://universaldependencies.org/u/pos/DET.html, https://universaldependencies.org/u/overview/morphology.html#pronominal-words.
15 The corpora were the following components from the UD collection Version 2.5: Chinese GSD, Dutch Alpino and

LassySmall, English EWT and GUM, French GSD, German GSD, and Indonesian GSD. In Indonesian, an additional check was performed due to somewhat idiosyncratic annotation choices, e.g. *akan*, which is a future marker meaning 'will, going to', and *dapat*, *bias*, and *mampu*, which express modality 'can, be able to' are annotated as adverbs. These words were excluded with the help of the tag 'M' in the language-specific POS set.