The Koala Part-of-Speech Tagset

Yvonne Adesam, Gerlof Bouma

University of Gothenburg
Språkbanken, Department of Swedish
{yvonne.adesam, gerlof.bouma}@gu.se

Abstract

We present the Koala part-of-speech tagset for written Swedish. The categorization takes the Swedish Academy Grammar (SAG) as its main starting point, to fit with the current descriptive view on Swedish grammar. We argue that neither SAG, as is, nor any of the existing part-of-speech tagsets meet our requirements for a broadly applicable categorization. Our proposal is outlined and compared to the other descriptions, and motivations for both the tagset as a whole as well as decisions about individual tags are discussed.

1 Introduction

Parts of speech, or word classes, are morpho-syntactic categories that divide the vocabulary of a language into groups to reflect a vocabulary item's morphological, syntactic, and semantic potential. While it is easy to show that words in general can be divided into different classes with the help of prototypical examples, it is not always easy to decide what class a particular word belongs to. There is no general agreement about such a classification, or even about the classification criteria. Part-of-speech tagging, analyzing words in running text according to a given part-of-speech inventory, is a common task when creating corpora. The part-of-speech inventory used for the task will depend heavily on which language is being analyzed, but also on factors like medium or genre, and on the purpose of the annotation.

In this article, we present an overview of the Koala part-of-speech tagset, suitable for annotating written contemporary Swedish. The paper starts with a discussion of previous part-of-speech categorizations for Swedish and our motivations for creating the Koala tagset, in Section 2. Then, Section 3 discusses the design principles behind the Koala tagset. The Koala part-of-speech inventory itself is described in detail in Section 4, with reference to other proposals for contemporary Swedish.[1] Section 5 highlights three words or groups of words which we treat differently from other classifications and which therefore need special mention. Finally, Section 6, goes briefly into our treatment of multiword units and their relation to the parts of speech.

[1]Parts of this comparative description have previously been presented at Svenskans beskrivning 2016 and the Swedish Language Technology Conference (SLTC) 2018.

2 Developing a New Part-of-Speech Tagset

The Koala part-of-speech tagset is part of a set of linguistic annotation guidelines developed for annotating written contemporary Swedish. The larger context also contains annotation layers with word senses and syntactic structure, in the form of head-marked constituents and grammatical functions (Adesam et al., 2015a). These layers are closely linked. Although this paper only details the part-of-speech layer, we will will also refer to the analysis at other layers when this link affects the part-of-speech categorization. The guidelines were developed within the four year project Koala, funded by Riksbankens Jubileumsfond. The Koala project aimed to improve the quality and relevance of the linguistic annotation in the corpus processing pipeline of Språkbanken Text, the Swedish language bank. The Koala guidelines have been used to annotate the 100 000 token Eukalyptus treebank[2] (Adesam et al., 2018). This treebank contains five different types of contemporary Swedish public domain texts, including texts which have not been professionally edited, such as blogs. The annotation will be applied to the corpora available through Språkbanken's corpus infrastructure Korp[3] (Borin et al., 2016, 2012).

Creating a new part-of-speech tagset is no small endeavour. However, the Koala tagset did not appear in a vacuum — it is of course based on existing grammatical descriptions, and has to fit into an existing language-specific field and tradition. In the following, we therefore discuss some important background considerations in creating the Koala part-of-speech tagset. We begin with the question of who is interested in parts of speech, and, related to this, who our intended audience is. We also review existing tagset proposals. From this, we move on to a discussion of our requirements for a categorization and why previous approaches do not meet these requirements.

2.1 Why Are Parts of Speech Interesting?

When presenting a new part-of-speech categorization, one may ask why and for whom parts of speech are relevant. One group interested in parts of speech are morphologists and syntacticians, although among these, functionalists may be more interested than formalists (see also Baker and Croft, 2017). Apart from having an interest in such categorizations, they are also potential users of the language resources we produce. To cater to this group, it is important that our choices are in line with (or at least not too different from) the current view on grammatical descriptions of the language. Both what we say and how we say it is important.

In typology, parts of speech are discussed in terms of their distribution across languages (for example, Haspelmath, 2012; Vogel and Comrie, 2000). The Koala tagset is not specifically geared towards typologists, as it does not define its categories in terms of cross-linguistic traits, but focuses on describing the Swedish language in particular. This can be contrasted with the Universal Dependencies project's (Nivre, 2014) explicit goal of being "good for linguistic typology"[4], which includes positing a cross-linguistic inventory of part-of-speech labels. (See Croft et al., 2017, and Osborne and Gerdes, 2019, for a critical discussion of Universal Dependencies and typology.) We will return to the typological

[2] `https://spraakbanken.gu.se/eng/resource/eukalyptus`

[3] See `http://spraakbanken.gu.se/korp`

[4] From `https://universaldependencies.org/introduction.html`; consulted 30 March 2019.

literature in Section 3, when discussing the design principles behind the Koala part-of-speech tagset, as it contains many relevant discussions about the basis for distinguishing between categories.

One would think that lexicographers would be interested in the details of a part-of-speech inventory, as this information is typically part of a dictionary entry. However, somewhat surprisingly, parts of speech seem only to receive passing mentions in lexicographic literature. Neither the *Routledge Handbook of Lexicography* (Fuertes-Olivera, 2017), nor the *Oxford Handbook of Lexicography* (Durkin, 2015) specifically talk about parts of speech. The *Bloomsbury companion to lexicography* defines part of speech as "the syntactic classification or grammatical role of a sense or entry," (Jackson, 2013, Glossary, p. 401) but does not discuss how this is determined or how such a classification should be done. The Swedish *Handbok i lexikografi* (Svensén, 2004) devotes a short chapter to the conventions for indicating part of speech, and notes that parts of speech are rarely treated in the meta-lexicographic literature. We therefore conclude that while parts of speech are of lexicographic interest, as they are part of the dictionary entry, there currently does not seem to be a thorough discussion about the choice of categorization, or the grounds for such a categorization.

Within the field of computational linguistics or language technology, parts of speech are considered a basic unit of annotation. In the standard pipeline model, they are the target output in the (word level) task of part-of-speech tagging, and necessary input building blocks for syntactic structure and other types of annotation. From this perspective, it is natural to assess a part-of-speech inventory in terms of how hard it is to apply, and how useful it is for downstream applications. The linguistic model implicit in the inventory is not as important. A new tagset may thus not be of general interest to language technologists, and may even be perceived as just another standard that one has to relate to — a nuisance. However, in our view, there is no reason to ignore linguistic knowledge for computational use. Creating a language resource, in our professional context, serves (at least) two purposes: The first is to create resources for language technology use, to serve as learning or evaluation material. The second is to make large scale text collections available for research, primarily linguistic or linguistically informed research. Since the primary user groups here are linguists and philologists, we consider it essential that the annotation is linguistically relevant and in line with the currently standard view on Swedish grammar.

2.2 Existing Part-of-Speech Categorizations for Swedish

The part-of-speech tagset of the Stockholm-Umeå corpus (SUC; Ejerhed et al., 1992) has been the predominant tagset for automatic part-of-speech tagging of Swedish text in the past two decades. Part of its success has been the one million word Stockholm-Umeå corpus itself, annotated with manually checked part-of-speech tags, morphological tags, lemmata, and name classes. Unfortunately, the corpus is released under a restrictive license, and is thus not freely available for, in particular, use in commercial applications. The billion word contemporary corpora available through Språkbanken have been automatically annotated with the SUC tagset. A number of slightly modified versions of the tagset have been used through the years, for example in the Granska tagger (Granska TextAnalysator, Knutsson et al., 2003), see also Carlberger and Kann (1999) and Forsbom (2008), and in the Swedish part of the parallel treebank Smultron (Volk et al., 2010).

Sweden has a long tradition of developing corpora and treebanks, which includes the pioneering work on Talbanken in the 1970s. The associated guidelines (MAMBA; Teleman, 1974) specify annotation at lexical (part of speech, morphological information) and structural (phrases, grammatical functions) levels. The MAMBA-based annotations have also been reused in later resources (Nivre et al., 2006, 2008; de Marneffe et al., 2014).

Both the MAMBA and the SUC guidelines predate the Swedish language reference grammar *Svenska Akadamiens grammatik* (SAG; Teleman et al., 1999). SAG contains the most thorough modern description of the language, and is the main point of reference for researchers and students of descriptive/theoretical Swedish linguistics. Therefore it is also our entry point in developing the part-of-speech tagset.

More recently, the work on the Universal Dependency project (Nivre et al., 2016) has been influential within the field of language technology, and we thus also need to consider the Swedish implementation of the Universal Dependency part-of-speech tagset (UDP), as announced in Nivre (2014) and described in the guidelines on the Universal Dependencies website.[5]

We are aware that there are many other descriptions of Swedish parts of speech, for example in the form of university level text books (see Josefsson, 2005, for an overview), which we do not review here. MAMBA, SUC and UDP have been used to develop considerable annotated resources. These, together with our baseline SAG, therefore form the background and point of comparison for our discussion.

2.3 Motivations for a New Tagset

The Koala part-of-speech inventory, and the syntactic and lexical semantic annotation models surrounding it, were developed with the aim of providing a description of contemporary written Swedish that is suited for different types of texts available through Språkbanken, and which aligns with the expectations of the linguistically informed users of these materials. The language we target is the written Swedish of the latter half of the 20th century and onwards. Works authored by non-professional writers nowadays are a salient part of the publically available body of text, for instance, in the form of blogs and social media texts. We therefore also want to cater for (productive) phenomena particular to such texts. The categorization does not handle phenomena specific to historical language varieties, for instance, the richer morphology of earlier stages of Swedish (see also the discussion in SAG I, 10–12[6]).

The complete annotation model was developed based on several principles. First, annotated information should be seen as an integrated whole, where the part-of-speech, phrase, and function categories have complementary roles. Secondly, we wanted a syntactic annotation schema which is easy to convert into other formalisms. This does not directly influence the part-of-speech tagset, but is still an important factor for the annotation at a general level. Thirdly, the distinctions should be easy for the annotators to understand and

[5]See `http://universaldependencies.org/guidelines.html`, and `http://universaldependencies.org/sv/` for the Swedish addenda; consulted 30 March 2019.

[6]SAG consists of four books, each part with its own page numbering. Chapters, from book two onwards, are numbered consecutively across books. When referring to particular locations in SAG, we either use Arabic numerals to refer to chapters in conjunction with a section number, or roman numerals I–IV to refer to books in conjunction with a page number.

to annotate, while still being linguistically motivated. Finally, we wanted an annotation schema in line with a modern view on descriptive Swedish grammar, which is why we use SAG as our point of departure.

There are considerable differences between the grammatical description in SAG and the other, more computational, descriptions intended for annotating natural language. These are not just differences in inventory and classification, but also in purpose and nature. While MAMBA, SUC and UDP are supposed to be an applicable set of annotation guidelines, SAG is a comprehensive grammar, which records all kinds of variations and subtleties of the Swedish language. Operationalization of the part-of-speech inventory is not a priority in SAG, and many ambiguities and delimitation problems are noted without being resolved. Although such a view is acceptable and necessary when discussing grammatical issues, this makes the description problematic for part-of-speech tagging, where clear-cut categories are needed for consistency and quality.

In addition, although SAG aims to be descriptive rather than prescriptive, it focuses on the written *standard* language (SAG I, 17–18, 29–30). The Koala guidelines are meant to be applied to a much broader range of material. In particular social media is a large part of the data we work with. When annotating authentic language data, we need to provide an analysis also for non-standard language, such as exploratory language use, language produced by language learners, language variants, and even errors.[7]

Using SAG directly is thus not an option. At the same time, we chose not to use SUC or MAMBA, since they predate and diverge from the descriptions in SAG. In addition, the previously well-used SUC tagset is only a lexical morpho-syntactic description, and does not describe the syntactic structure of the whole sentence. Finally, SUC and MAMBA also define part-of-speech categories which are very fine-grained, while we have striven for a tagset that is minimal, but with most relevant distinctions present.

When we started developing the guidelines for the Eukalyptus treebank, the Universal Dependencies framework for Swedish was still in its infancy. Although it now is a possible choice for developing a Swedish treebank, it still diverges from the description in SAG, and some categories are clearly influenced by the cross-linguistic ambitions of the Universal Dependency programme. We also note that as a system, Universal Dependencies is a form of dependency grammar, and prefers heads to be content words, rather than function words. In our view, phrase structure analysis with phrases projected from either content or function words enjoys a wider acceptance in the Swedish descriptive linguistic community.

Taken together, these considerations led us to develop a new SAG-influenced annotation model, of which the part-of-speech inventory described in this paper forms the morpho-syntactic layer.

3 Design Principles of the Koala Tagset

Through history, there have been large variations in what part-of-speech categories are used, as well as in the number of categories. One of the most important reasons for these differences is disagreement over the criteria for dividing words into groups. Part-of-speech categories are an encoding of the human knowledge about words and proficiency in using a vocabulary — a surrogate representation (Davis et al., 1993) which by necessity is flawed

[7]That is, we wish to annotate varied language, but not necessarily the variation itself.

or partial. Using different criteria, or weighting the same criteria differently, will result in different encodings. It is obvious that a focus on form will give a different description than a focus on function. Trask (1999) identifies at least four different language-internal criteria: meaning, distribution, inflection, and derivation. Josefsson (2005) also discusses cross-linguistic criteria. We should also add tradition as a potential criterion, but see Pullum (2009), who criticizes putting too much weight on such arguments. As pointed out in Haspelmath (2015), there are also two different ways of looking at categorization criteria: as diagnostic tests for an a priori established inventory; or as defining criteria, "necessary and sufficient" to determine categories.

The part-of-speech definitions of the Koala tagset are based on morphology and syntactic distribution, and to a lesser extent semantics. We prefer a categorization where these three criteria converge. However, where this is not the case, we use a word's inflectional properties as the primary criterion for determining its part of speech — inflectional contrasts are typically easily observed. We go by a word's syntactic distributional properties, when the signal from inflection is non-decisive or when the distributional facts are particularly clear. We do not use semantics as a systematic criterion. Semantics is, however, a factor in a word's syntactic distribution and also plays a role in deciding to which lemma(ta) a form belongs, and therefore influences our perception of this distribution.

Distributional evidence needs some special attention, since we developed the part-of-speech categorization in tandem with the syntactic annotation model (Adesam et al., 2015a). We have tried to minimize the overlap between the two levels with regards to which generalizations they capture. For instance, the tendency of a verb particle to appear between its verb and object could be modeled by appealing to its part of speech, or by appealing to its grammatical function. All else being equal, we prefer to use grammatical function to capture such a generalization, and therefore, this distributional fact should not be used to determine a verb particle's part of speech. Thus, verb particles may belong to one of many parts of speech, although they typically are prepositions or adverbs (*hoppa i*, 'jump into', lit. 'jump in'; *ta hit*, 'bring', lit. 'take here'). Other examples are head-like adjectives in noun phrases (*de anställda* 'the employed'), which we categorize as adjectives and not nouns; or predicatively used interjections, which are just that and not adjectives (*vara blää* 'to be yuck').

Cross-linguistic criteria may be suitable for pedagogic reasons in grammar books for students, or for typological reasons. Typological reasons lie at the basis of UDP's approach, which starts with a set number of categories to choose from when describing a particular language.[8] We do not introduce categories merely because they are present in other languages, since our focus is the Swedish language. Recasting Haspelmath's diagnosing-defining dichotomy as a continuum, we can put UDP's approach more towards the diagnosing side, whereas we try to stay on the defining end of the spectrum, even though we admit that, sometimes, diagnosing is the best we can do.

We now turn to the specifics of the Koala part-of-speech inventory. It consists of 13 coarse-grained categories. Eleven of these correspond to traditional parts of speech, while two (foreign material and symbols) are necessary to annotate authentic language. The former eleven are closely related to the 13 categories of SAG. We have, however, removed two categories which we found were not distinctive enough, and we have redrawn some

[8]It should, however, be noted that UDP does not claim that this set of categories is cognitively universal, just that this set is enough for describing language in terms of parts of speech.

Table 1: Koala parts of speech and features.

Part of Speech	Features
Adjective	degree, gender, number, definiteness, relative
Adverb	degree, relative
Common noun	gender, number, definiteness
Coordinator	
Foreign word	
Interjection	
Numeral	
Preposition	
Pronoun	gender, number, definiteness, form, relative
Proper name	
Subordinator	
Symbol	type
Verb	mood/finiteness, voice, tense
Any part of speech	genitive, abbreviation, partial ellipsis of compound

category borders where SAG already in its discussion opens for this possibility. These deviations will be discussed in Section 4.

The 13 part-of-speech categories are supplemented by a set of features, for information like agreement properties, verbal tense or finiteness, whether a word is an interrogative, whether is has a genitive marker, etc. Mostly, these features relate to inflection, that is, the value of a feature changes as a word's inflection is varied. In fewer cases, features represent inherent properties of a lexical item (for instance, noun gender or definiteness of a pronoun).

Overall, the level of abstraction is similar to that of SAG (at the level of part of speech) and SUC (at the level of morphology). Our morphological features are meant to primarily encode morphologically marked — that is 'visible' — information, and we therefore do not include tags for information such as verb valency or the count/non-count distinction in nouns. We can, however, envision future extensions of the tagset to include more such information, or other information specific to certain types of corpora, such as dialogue or historical inflection. Despite our primary focus on marking, we include noun gender and pronominal definiteness, as they form part of a cluster of properties that apply across the categories noun, adjective and pronoun.

Most of our features are restricted to certain parts of speech, and we refer to these as *specific features*. In the next section, they will be discussed together with the individual part of speech they apply to. In addition, we also use three *general features*, which may be added to items of any part of speech. These general features will be discussed separately in Section 4.1. An overview of the categories and the features is given in Table 1. For any token (text word), one part-of-speech tag is chosen. A type (lexical item) may be

associated with several parts of speech. A specific feature can be unambiguous (gender: N; neuter), associated with syncretism (number: SG|PL), or unspecified (degree: -).

Our treatment of syncretism (or generally non-inflecting forms) depends on how systematic it is. Syncretism that is predictable given a combination of feature values triggers predefined disjunctive feature levels. For instance, plural adjectives are systematically invariant for gender, so items annotated with 'number: PL' are automatically given 'gender: C|N' (common or neuter), independent of context. Likewise, adjectives in the comparative do not inflect at all, which means that their definiteness, number, and gender agreement features are all set to disjunctive values. However, if the syncretism is not predictable from the set of features, items are annotated fully specified according to context. For instance, an indefinite neuter noun like *hus* 'house' may be singular or plural, however, since this is not the case for all neuter indefinite nouns, we infer number from the syntactic context or even from the discourse.

Not all levels of a feature need to be specified. For example, a verb is unspecified for tense ('tense: -') when used in the imperative ('mood: IMP'). Some of the features have implicit default values, that apply when a feature is left unset. For instance, if the feature that marks abbreviations is not set, the default implication is that a word is not an abbreviation.

An important issue when handling text is segmentation, that is, defining the units of annotation. SUC follows a rather strict schema where space separates tokens (although abbreviations such as *t ex* 'e. g.' are processed as *t_ex*). This is also a common approach within the language technology community, since it is deterministic and transparent. From a linguistic point of view, however, space is an imperfect token delimiter. In the Koala schema, we allow tokens to contain spaces. For example, both *idag* and *i dag* 'today' are one token, and so is *Mont Blanc-tunneln* 'Mont Blanc tunnel.DEF'. In some cases, it may also be necessary to split what is written as one word into multiple tokens. For example, non-standard orthography *serunte* 'don't you see' is segmented into the substrings *ser u nte*, lit. 'see you not'. In contrast to the treatment of clitics and fused forms in UDP, we do not replace the segments with full, standard orthography word forms in such cases, which would have been *ser du inte*, as the information this would add can be inferred from the lemmatization.

Allowing tokens to contain spaces may be a problem for automatic processing, as they introduce segmentation ambiguity. Some of these cases can be properly analyzed as multiword units (Sag et al.'s, 2002, "words with spaces"). Following Borin et al. (2013), we annotate multiword units with part-of-speech labels, but as phrases on the syntactic level (Adesam et al., 2015b). They therefore reside on a level between the tokens and the phrases in the annotation. This moves the identification of groups of tokens to the syntactic annotation level, where such problems fit naturally. Multiword units are discussed further in Section 6.

4 The Koala Inventory

We start our overview of the Koala inventory by discussing the general features, which may apply to any part of speech. The rest of the section deals with the individual parts of speech and their related specific features. We begin with the nominal categories, those

where the dimensions of definiteness, gender and number are relevant: common nouns, proper names, adjectives, and pronouns. We then consider verbs, which inflect inter alia for tense and finiteness, after which we discuss parts of speech with no or little inflection: adverbs, prepositions, and subordinators, as well as coordinators, numerals, and interjections. We end our overview of the inventory with two categories that are not parts of speech in the traditional sense: symbols and foreign material.

The purpose of each section describing a part-of-speech category is to give a brief characterization in terms of inflection and distribution. The sections also contain some motivation for the choices made and indicate in which way we deviate from other proposals and analyses. However, the part-of-speech descriptions below only give a very general and, by necessity, deficient view of linguistic reality. They are not meant to be taken as attempts to write a language description — we refer to SAG for this — nor as annotation manuals, which would contain less motivation, but more examples and arbitration for hard or ambiguous cases.

The labels used for examples throughout this article differ from those used in the annotated corpus, to make the examples easier to read. The full tagset with part-of-speech and feature labels is listed in the Appendix.

4.1 General Features

Table 2: General features

Abbreviation	(default: no), yes
Partial ellipsis	(default: no), yes
Genitive	(default: no), yes

The features which may apply to items of any part of speech are listed in Table 2: abbreviation, partial ellipsis, and genitive. All three are binary features. The feature abbreviation is set for abbreviations, in addition to any annotation that follows from their regular word class. Abbreviations can be written with or without periods: *ung*, *ung.* 'approximately.ABBR' (full: *ungefär*), which is an adverb; and with or without spaces, for example, *bla*, *bl a*, *bl.a.*, *bl. a.* 'amongst other things.ABBR' (*bland annat*), also an adverb — but see Section 6 on parts of speech for multiword units for the written-out form.

Ellipsis below the word level in a coordination is marked with the partial ellipsis feature. We analyze such partial words as if they were complete. So in *vit- eller svartpeppar* 'white or black pepper', we have the common gender noun *vit-* 'white pepper.PARTIAL'. Likewise, *lägenhetssäljare och -köpare* 'apartment seller and buyer' has *-köpare* 'apartment buyer.PARTIAL'. SUC has a similar feature, however, in contrast to our approach, SUC annotates for just the realized part, which would make *vit-* 'white' an adjective. The advantage of our approach is that we are not forced to annotate bound morphemes according to our part-of-speech inventory. So, there is no need to decide whether the first conjunct in *tvätt- eller diskmaskin* 'washing or dishwashing machine' is from the verb *tvätta* 'to wash' or the noun *tvätt* 'laundry', as we annotate it according to the implicit *tvättmaskin* 'washing machine'. Similarly, to us, *be- och omarbeta* 'manipulate and change'

(lit.: '*be-* and rework') is just a conjunction of verbs and *o- och underbetalda* 'un- and underpaid' a conjunction of adjectives.[9]

The final of the three general features marks the genitive. Historical stages of Swedish had a four case nominal inflection system, which included the genitive case. In descriptions of Swedish, including SAG, it is common to find the genitive given as the only remaining case suffix, appearing invariably as *-s* on the nominal categories. However, much like English, Swedish has a 'group genitive', where the *-s* marking appears on the right edge of a noun phrase (SAG 14 §79). We therefore find *-s* marked words (Examples 1abcd) of almost any kind — as long as they appear noun phrase final, including in noun phrases with relative clauses — even though it is not as frequent as *-s* marking on nominal parts of speech and even though it is associated with less formal registers. We therefore use a general feature for the genitive, to mark the presence of a genitive *-s*, and remove the specific case features from the nominal categories.[10]

(1) (a) huset-s tak
house(N).SG.DEF-GEN roof
'roof of the house'

(b) karln där-s fru (s)[11]
the guy there.GEN wife
'that guy over there's wife'

(c) [D]et är den som talar-s egen åsikt [...]. (a)
it is the one REL speaks.IND.PRS-GEN own opinion
'It is the opinion of the person who speaks.'

(d) nån som jag var arg på-s varuvagn (a)
someone REL I was angry on-GEN trolley
'someone I was angry with's trolley'

There are also examples of genitive marking on a phrase's head, rather than its right edge: The formal register (2a) is an example where a postnominal preposition phrase directly follows its head noun, example (2b) shows a possessive noun phrase with an extraposed postmodifier (both examples from Börjars, 2003, to which we refer for extensive discussion). We apply the GEN label in these cases, too.

(2) (a) enskilda individer-s vid Operan yrkesskicklighet (a)
single individual(C).PL.IND-GEN at the Opera professional skill
'the professional skill of separate individuals at the Opera'

[9]The latter two examples show the phenomenon of partial ellipsis in its extreme form. Although they are attested in edited magazine and newspaper text, we acknowledge that such cases are relatively infrequent. The indeterminacy of the first part in a compound, such as *tvätt-*, is more common, however.

[10]SAG's position in this matter is slightly curious. They mention the possibility of attaching *-s* to other parts of speech in several places (in the glossary entry for *genitive*, in SAG 2 §76 note 1, and in SAG 14 §79, to name just three) and consider genitive marking to be a marking of the phrase. Still, they explicitly introduce genitive as case on the nominal parts of speech, without discussing whether this is an appropriate way of characterizing the suffix.

[11]The examples in the rest of the paper are either constructed (no marking), attested, found in Korp or online (a), or examples taken from SAG (s).

(b) fotbollssupportrarna-s skrik som sett sitt lag förlora
football supporter(C).PL.DEF-GEN shouts REL seen their team lose
'the shouts of the football supporters who had just seen their team lose'

By using a general feature, although marked on words, we remain agnostic about the morphosyntactic status of the genitive, or even whether it is a unitary phenomenon. This is also why we do not segment genitive *-s* into its own token, as done, for example, in the English Penn treebank (Marcus et al., 1993).

SUC, SAG, and UDP all have the nominative vs genitive case distinction on the nominal parts of speech. MAMBA holds a middle ground as it has a feature 'genitive suffix', which, however, only applies to certain parts of speech.

4.2 Common Nouns

Table 3: Common noun-specific features

Number	singular, plural
Definiteness	indefinite, definite
Gender	common gender, neuter

The most pervasive and obvious inflectional property of Swedish common nouns is that they allow definiteness inflection, as in *ett hus* 'a house.IND', *det hus-et* 'that house-DEF' and *två bilar* 'two cars', *de bilar-na* 'those cars-DEF'. Common nouns also have number, which in many cases can be systematically varied: *lamp-a* 'light-SG', *lamp-or* 'light-PL'. Finally, nouns are inherently specified for gender, which is most clearly reflected in the agreement constraints imposed on accompanying material such as determiners: *en sak* 'a.C thing(C)', *ett ting* 'a.N thing(N)'. Gender also factors into generalizations about an item's inflectional paradigm. Syncretism of number is common for indefinite neuter nouns: *ett hus* 'one house(N).SG.IND', *två hus* 'two house(N).PL.IND'. Syncretism of definiteness marking is rare, although it is the rule in a group of deverbal nouns ending in *-an*: *en längtan* 'a longing.IND', *den där längtan* 'that longing.DEF'. Since these often do not pluralize (or require suppletion), argumentation for their status as nouns needs to depend completely on evidence from distribution. The common-noun specific features are summarized in Table 3.

A noun phrase, which may be a bare noun, can generally appear as argument of a verb or preposition (3a), and as predicate with for instance copular *vara* 'be' (3b). When marked with genitive *-s*, noun phrases can appear as determiners to other nouns (3b). Headed by certain nouns, noun phrases can function as time or frequency adverbials (3c). When the noun phrase also contains determiners and adjectival attributes, there has to be agreement with respect to gender, number, and definiteness.

(3) (a) Kvinnan dricker kaffe med mjölk.
woman(C).SG.DEF drinks coffee(N).SG.IND with milk(C).SG.IND
'The woman is drinking coffee with cream.'

(b) Hon är läkare / året-s nobelpristagare.
she is doctor(C).SG.IND year(N).SG.DEF-GEN Nobel laureate
'She is a medical doctor / this year's Nobel prize winner.'

(c) Jag var där tio gånger.
I was there ten time(C).PL.IND
'I was there ten times.'

Word mentions (that is, meta-linguistic use) are very restricted in their inflection and inherent properties. We consider mentioned material to be nouns, irrespective of their would-be part of speech when used. Our motivation for this is their syntactic distribution. In particular, they may show up as heads in noun phrases (4).

(4) Det där lågt viskade "dra!" gav mig kalla kårar.
that softly wispered leave(N).SG.IND gave me cold shivers
'That softly wispered "leave!" gave me the willies.'

MAMBA specifies a number of finer categories, such as deadjectival and deverbal nouns, and also one for 'metanouns'. UDP annotates meta-linguistic uses according to their original category. We find both our approach and UDP's to have advantages and disadvantages. If a meta-linguistic token is annotated with its original category, one might end up with an uncommented dissonance, such as a verb heading a noun phrase. In the wider picture of our full annotation schema, which includes phrasal syntax, this would give us additional trouble, since we enforce strict constraints on the relation between the type of a phrase and the category of its head. On the other hand, from the perspective of the dictionary or computational lexicon, it is unfortunate to have just about anything potentially tagged as a noun. And a corpus user wishing to extract a list of nouns would probably also be surprised to see, for example, *dra* from (4) listed as a noun. For the future, a solution that combines both these sides is worth investigating: an indication of noun-like behaviour and an annotation of the original category. In this respect, MAMBA's special (sub)category of metanouns is a step in the right direction, but note that there is no indication of the original part of speech there.

In other aspects, UDP's and SUC's annotation for nouns is practically the same as ours. MAMBA's noun category, in contrast, includes both common and proper nouns, but does not specify inherent morphological features such as number and gender.

4.3 Proper Names

Proper names have no specific features. From a distributional perspective, proper names share many characteristics with common nouns. In particular, they behave much like definite (count) nouns, which may occur without attributes. However, their inflectional behaviour is quite distinct. Gender and number are not visibly marked for proper names, and they do not take definiteness marking in contexts that require it for common nouns (5ab). They may be accompanied by attributive modifiers in a noun phrase.

(5) (a) Det vackra Nora i Bergslagen
the beautiful Nora in Bergslagen
'beautiful Nora in the Bergslagen-district'

(b) Halva Göteborg / *stad / staden
half Gothenburg city(C).SG.IND city(C).SG.DEF
'half of Gothenburg / the city'

(c) ett tomt Gamlestaden
a(IND).SG.N empty.POS.SG.IND.N Gamlestaden
'Gamlestaden, which was empty'

Although a proper name may appear to have definite morphology, such as *stad-en* 'city(C).SG-DEF' in *Gamlestaden*, this apparent morphology does not have to agree with the context. This is illustrated in (5c), where *Gamlestaden* heads a neuter, indefinite noun phrase. SAG 3 § 15–16 discusses the agreement behaviour of proper names.

There are a number of borderline cases to consider. First, in (6a), in a plural context, a person name (optionally plural marked) is used to denote bearers of that name rather than an individual, and in (6b), the definite marked manufacturer's name denotes a vehicle produced by them.

(6) (a) Blåvitt hade tre Glenn / Glenn-ar
Blue-white had fyra Glenn(C).PL.IND / Glenn(C)-PL.IND
'IFK Göteborg had three players named Glenn'

(b) Vi bilar ner i volvo-n.
we drive down in Volvo(C)-SG.DEF
'We'll drive south in the Volvo.'

To accommodate the increased inflectional freedom in (6), but also because the underlined nouns do no longer *name* their referents, we do not consider them to be proper names, and treat them as common nouns instead.

Second, within a context, it is not uncommon to find the definite (singular) of a descriptively adequate common noun used in a name-like fashion — sometimes marked in writing by capitalization: *Riksdagen* 'parliament.DEF' (to mean 'The Riksdag of Sweden'), *Kungen* 'king.DEF' ('Carl XVI Gustaf'), *chefen* 'boss.DEF'. As a rule of thumb, we analyze these as common nouns, and reserve the proper name label for less ambiguous cases.

Finally, titles of works form a particular subset of proper names. In (7), *Upp*, which would otherwise be an adverb, may head a noun phrase by virtue of being the name of a work.

(7) Pixar studios Upp premiärvisades i Cannes.
Pixar studio's Up premiered in Cannes
'Pixar studio's Up premiered at Cannes.'

We consider *Upp* (just like *Pixar studios* and *Cannes*) to be a proper name. The same considerations as for the annotation of meta-linguistic material apply here as well. For a generalization of our approach to multiword titles, see Section 6.

MAMBA, as mentioned, considers names to be part of the noun category. They also differentiate between person names and other proper nouns, as they mark person and non-person for all nouns.

4.4 Adjectives

Table 4: Adjective-specific features

Degree	positive, comparative, superlative
Number	singular, plural
Definiteness	indefinite, definite
Gender	common gender, neuter, masculine
Interrogative/Relative	(default: no), yes

Table 4 lists the adjective-specific features. Adjectival inflection in part follows the same dimensions as common noun inflection, but notably with different suffixes. There is a systematic syncretism of all definite and all plural forms. In addition, the definite singular may optionally be realized by *-e*, instead of the syncretic *-a*, to mark, mainly, a male or generic human referent (SAG 4 § 68).[12]

Some adjectives show defective nominal inflection paradigms and/or lack of inflection. Adjectives can have morphological comparison: *tuff–tuffare–tuffast* 'tough/tougher/toughest', or periphrastic comparison: *mer/mest utmanande* 'more/most challenging'. We only use the labels comparative/superlative for the former. Comparatives do not inflect any further, superlatives only inflect for definiteness (SAG 2 §63).

In Swedish, interrogative and relative forms partially overlap, which is why we, like SUC, mark them with one binary feature, here abbreviated IR. We only find one adjective marked as such, the interrogative *hurdan* 'what kind', counterpart of *sådan* 'such': *ett hurdan-t hus*, lit. 'a what-kind(IR) house', more common in Finland Swedish.

A characteristic of phrases headed by adjectives is that they may occur as prenominal attributes (8a) or predicates (8b) in a large range of predicate constructions: for instance with *vara* 'be' and *bli* 'become/turn into/grow/go', with verbs like *verka* or *se . . . ut*, both 'seem/look', and as secondary predicates.

(8) (a) den tuffa banan
the(DEF).SG.N difficult.SG.DEF.C|N track(C).SG.DEF
'the difficult track'

(b) Banan såg tuff ut
track(C).SG.DEF looked difficult.SG.IND.C VPRT
'The track looked difficult'

Adjectives agree in number and gender, but they take definite marking only when they are prenominal. Some adjectives only occur predicatively or attributively. Modifying an adjective requires an adverbial.

In the singular indefinite neuter, an adjective can typically function adverbially (9).

[12]We mark these with the gender value masculine, but wish to point out that contemporary standard Swedish does not have a three-way syntactic gender in the same way as, for instance, German does.

(9) Hon sjunger hög-t.
She sings loud-SG.IND.N
'She sings loudly.'

We follow SAG in still treating these as adjectives. Of the other descriptions considered here, SUC is alone in treating these as deadjectival adverbs.

It is possible to form a nounless noun phrase in which the adjective plays a head-like role. We can identify these as adjectives on the basis of the actual suffix (SAG 2 § 70), and the fact that they, unlike nouns, allow adverbial pre-modifiers (10).

(10) De mycket sjuk-a har det tufft i kylan.
the(DEF).PL.C|N very ill-PL.IND|DEF.C|N have it difficult in the cold
'It's difficult in the cold for the very ill.'

Present participles are treated as adjectives or nouns. Here, too, the distinction is made on the basis of inflection (definiteness marking) and modification. See Section 4.6 regarding participles.

4.5 Pronouns

Table 5: Pronoun-specific features

Form	(default: n/a), subject, object, possessive
Number	singular, plural
Definiteness	indefinite, definite
Gender	common gender, neuter, masculine
Interrogative/Relative	(default: no), yes

With respect to inflection, we can observe three clusters of pronouns: those with adjectival inflection for number and gender (*ingen*: 11a), those that do not inflect at all (*varje* 'each', *allting* 'everything'), and the personal pronouns, which have subject, object and possessive forms (*man*: 11b). The possessives of some of the members of the third cluster are themselves in the first cluster (*jag*: 11c; but *min*: 11d).

(11) (a) ingen – inget – inga
SG.C SG.N PL.C|N
'no/none/noone'

(b) man – en – ens
SUB OBJ POSS
'one/people'

(c) jag – mig
SUB OBJ
'I/me'

(d) min – mitt – mina
POSS.SG.C POSS.SG.N POSS.PL.C|N
'my'

As can be seen in (11d), we mark agreement features according to the constraints imposed on a containing noun phrase, not according to properties of the referent. For all pronouns,

definiteness is considered an inherent feature. Indeed, for several of the pronouns, marking definiteness can be considered their main function. Interrogative/relative pronouns can be found among both personal pronouns (*vem*, *vars*, *vad*: 12abc) and adjectivally inflected pronouns (*vilken*: 12d). The contrast *vems* (12a) vs *vars* (12b) shows that not all interrogative and relative forms coincide.

(12) (a) vem – vems
(IR).SUB|OBJ (IR).POSS
'who/whose (interrogative)'

(b) vars
(IR).POSS
'whose (relative)'

(c) vad
(IR).SUB|OBJ
'what'

(d) vilken – vilket – vilka
(IR).SG.C (IR).SG.N (IR).PL.C|N
'which'

Table 5 combines the features that apply to all of these three clusters of pronouns.

The fact that we separate the general genitive from the pronoun specific possessive, means that we can naturally handle cases of double marking (13a) and cases of genitive marking on an embedded personal pronoun (13b).

(13) (a) Cykeln är bror min-s. (s)
the bike is brother(C).IND.SG my(DEF).POSS.SG.C-GEN
'The bike is my brother's.'

(b) en vän till mig-s lillebror (a)
a friend to me(SG.DEF.C).OBJ-GEN little brother
'my friend's little brother'

Generally, pronouns can function on their own as a noun phrase. As far as other material in the noun phrase is concerned, we can distinguish two groups of pronouns. One group routinely functions as determiner, and freely combines with subsequent adjectival material. Three members of this group (*någonting* 'something', *ingenting* 'nothing', *allting* 'everything') preclude the presence of a head noun (14a), whereas others can co-occur with a noun (14b). Determiners may be combined in one noun phrase, within certain limits. The other group consists of pronouns used as heads. They do not combine with determiners, except for with pre-determiners (*halva jag* 'half of me', *båda oss* 'both of us'). They may occur, in a restricted way, with preceding adjectives (14c).

(14) (a) någonting grönt (*blad)
something(SG.IND.N) green leaf
'something green'

(b) något grönt (blad)
some(IND).SG.N green leaf
'something green' alt. 'some green leaf'

(c) hela (*den) underbara han
whole(DEF) the wonderful.SG.DEF.C|N he(DEF.SG.C).SUB
'all of him, who, by the way, is wonderful'

As should be clear from the description, the class of pronouns forms a rather heterogeneous group, and it is hard to pin down a distinguishing set of features. Compared to SAG, we have a smaller class of pronouns, treating as adjectives a set of pronouns that SAG calls relational pronouns and describes as very adjective-like (SAG 4 §71, 5 §196ff), such as *annan* 'other', *egen* 'own', or *sista* 'last'. SAG, SUC and MAMBA distinguish many more different types of pronouns. SUC and UDP have different categories for determiners and (non-dependent) pronouns — we consider this to be a difference in syntactic function. In our inventory, *en/ett* 'a/one' is a pronoun, irrespective of whether it has an indefinite article sense or a numeral sense. See Section 5.1 for motivation.

4.6 Verbs

Table 6: Verb-specific features

Mood/Finiteness	indicative, subjunctive, imperative, infinitive, supine
Tense	(default: n/a), present, past
s-Form	(default: no), yes

Swedish verbs inflect along dimensions of mood/finiteness, tense, and marking with a suffix *-s*, as summarized in Table 6. The most easily recognizable and pervasive property is the inflection for tense on the indicative, in a sentence pair like (15).

(15) Jag duscha-r idag / duscha-de igår.
I shower-IND.PRS today shower-IND.PST yesterday
'I shower today / showered yesterday.'

With the exception of certain archaic expressions, Swedish verbs do not agree, hence the lack of any verb-specific features taken from the nominal domain. The tense feature also applies to subjunctive forms, but not to the imperative, infinitive or supine levels of mood/finiteness. In the context of Swedish grammar, the term 'supine' refers to a distinct form whose primary use is in the periphrastic perfective, as in (16).

(16) Jag har redan duscha-t i dag.
I have already shower-SUP today
'I have already showered today.'

Inflectional paradigms may show defects in the tense and mood/finiteness dimensions, for instance for a number of irregular auxiliary verbs.

From a distributional point of view, indicative verbs are characterized by their ability to immediately follow a subject and form an affirmative root sentence. If the subject is a personal pronoun, it has to be in the subject form (17a). We do not distinguish subcategories of verbs based on valency. Some valency frames allow combination with another verb, which then has to be in the infinitive or supine (17bc).

(17) (a) Superhjältinnan / Hon / *Henne gäspar.
the superheroine she her yawn.IND.PRS
'The superhero / She yawns.'

(b) Hon försöker gäspa / *gäspar / *gäspat.
she try.IND.PRS yawn.INF yawn.IND.PRS yawn.SUP
'She tries to yawn.'

(c) Hon har *gäspa / *gäspar / gäspat.
she have.IND.PRS yawn.INF yawn.IND.PRS yawn.SUP
'She has yawned.'

The subjunctive is more rarely used in contemporary Swedish. Subjunctives distribute like indicatives, except that they express hypothetical states of affairs.

(18) Jag / *Mig vore ingenting utan er! (a)
I me be.SUB.PST nothing without you
'I would be nothing without you!'

The subjunctive *vore* 'be.SUB' is the only form occurring frequently (SAG 7 §41).

Related to valency is the distinction between auxiliary and main verbs. UDP marks auxiliaries, and views copulas as auxiliaries. We follow SAG's reasoning that the border between auxiliaries and main verbs is fuzzy, and do not distinguish auxiliaries from other verbs; SUC presents a similar reasoning. MAMBA does not have a single label for auxiliaries. They do, however, mark a number of verbs with their own singleton part-of-speech category, among which are some auxiliary verbs.

SAG, SUC, and MAMBA treat participles as their own category, whereas we, like UDP, view them as deverbal adjectives. Present participles do not inflect, but have an adjectival distribution, while perfect participles show both adjectival agreement inflection and adjectival distribution. Perfect participles are also used in the periphrastic passive, without consequences for their inflectional behaviour. Treating them as adjectives avoids having to add nominal inflection to the verb-specific features.

The suffix *-s* cross-cuts tense and mood/finiteness levels, and is found with a number of arity reduced realizations: passive (*ätas* 'be eaten'), reflexive/reciprocal (*träffas* 'meet/come together'), habitual/progressive aspect (*bitas* 'biting (repeatedly/habitually)'), all of which have transitive *s*-less counterparts. Some verbs only have forms with *-s* (*brås* 'take after/resemble', but **brå*), and for others the link between the two forms is not synchronically meaningful (*finna* 'find', *finnas* 'exist'). We apply the s-form feature to all of these, following SUC.

Passivization in Swedish touches upon all three issues of s-marking, auxiliaries, and participles, because of the existence of both a morphological and a periphrastic passive. As mentioned, the morphological passive coincides with other voice-related morphology. Furthermore, the periphrastic passive involves a participle (that is, an adjective) with regular adjectival agreement behaviour. The Swedish passive is therefore never morphologically distinct. As a result of our focus on explicit distinctions (see Section 3), we thus currently offer no way of identifying passives of either form. UDP and MAMBA mark the passive as an inflectional category. They also annotate the periphrastic passive distinctively.

4.7 Adverbs, Prepositions and Subordinators

Table 7: Adverb-specific features

Degree	(default: n/a), positive, comparative, superlative
Interrogative/Relative	(default: no), yes

Prepositions and subordinators do not have any specific features; the adverb-specific features are given in Table 7. A small group of adverbs allows degree inflection, which in form resembles the adjective degree suffixes (*ofta–oftare–oftast* 'often, more/most often'). Otherwise, adverbs, prepositions[13] and subordinators do not show any inflection. The adverbs contain a relatively large subset of interrogative and relative forms (*hur* 'how(IR)', *varpå* 'upon which(IR)').

Phrases headed by adverbs, prepositions, and a subset of the subordinators all have in common that they may be used as adverbials — as modifiers of verbs, adjectives, and members of their own groups. Several of them can also head oblique complements. What distinguishes the three parts of speech are the constraints on accompanying material: In general, adverbs do not require any such material (*länge* in 19a); prepositions combine with noun phrases (*på* and *i* in 19a) or marked subordinate clauses (19b); and subordinators must be accompanied by unmarked subordinate clauses. So, in (19b), the complement of the preposition *mot* must have *att*-marking, whereas in (19c) the complement of subordinator *bara* may not.

(19) (a) Jag har hejat på GFC länge / i många år.
I have cheered on GFC long for many years
'I have supported GFC for a long time / for many years.'

(b) Vilka vägar finns mot *(att) bli klimatneutrala? (a)
Which paths exist towards to become climate neutral
'Which paths lead towards becoming climate neutral?'

(c) Bara (*att) du kommer så ordnar jag frieriet. (a)
just that you come then arrange I marriage proposal
'As long as you get here, I'll arrange the marriage proposal.'

When used adnominally, phrases headed by adverbs, prepositions, and subordinators occur in peripheral positions: typically postnominally, but cases like focus adverbs may also appear in initial position (20).

(20) åtminstone det långa mötet igår
at least the long meeting yesterday
'at least the long meeting yesterday'

[13] One preposition can be said to show inflection of degree: *nära/närmare/närmast dig* 'close/closer/closest to you'.

Swedish has prepositions, postpositions and circumpositions — we gather all of these under the term preposition. Subordinators always precede their complements.

Three subordinators require special mention (see SAG 11 §7): *som* 'that/as', *än* 'than', and *att* 'that/to'. The subordinator *som* is among other things used as a relative clause marker. It can also, like *än*, be used in comparisons, and they then combine with a much wider array of complements than indicated above, for instance because of comparative deletion. We mark these uses as subordinators, without considering what a fully spelled out sentence would look like (SAG 26 §1, notes). See Section 5.2 for further discussion about *som*. Finally, *att* can be either of two homographic subordinators: one that combines with an unmarked finite clause, and one that combines with an unmarked infinitival clause/VP. These clauses, initiated by the subordinator, can be used adverbially or nominally.

SUC has a separate category for interrogative and relative adverbs, which we mark with a feature, as well as a category for verb particles, most of which we consider to be adverbs or prepositions (see Section 5.3). Only SUC groups adverbial adjectives with suffix *-t* among the adverbs (Section 4.4). UDP separates negations from adverbs and tags them as 'particles', while we treat them as any other adverb. MAMBA has a very fine-grained set of subcategories for different uses of adverbs and subordinators. Finally, our view that infinitival *att* is a subordinator differs from the other descriptions. It is considered its own category by SAG, SUC and MAMBA, and a particle by UDP.

4.8 Coordinators

Coordinators are non-inflecting words whose prototypical function is to combine two similar parts (conjuncts) into one whole of the same kind (the conjunction), without a hierarchical difference between the conjuncts. We emphasize that this description leaves underspecified the type of the conjuncts and in what sense they must be similar. Although prototypical cases have conjuncts that are of the same part of speech and/or phrasal category (example 21a, which combines two nominal conjuncts), conjunctions are not constrained to these (example 21b, where the predicative is a coordinated noun phrase and adjectival phrase).

(21) (a) barnen och jag
the kids and I
'the kids and me'

(b) Är du ekonom och duktig på projektledning? (a)
are you business administrator and good at project management
'Are you a business administrator and good at project management?'

Coordinators may have inherent restrictions on whether they allow more than two conjuncts, and which type of conjuncts they take. We do not capture such distinctions, neither through part of speech nor by means of features. There are no coordinator-specific features in our annotation model.

Coordinators may also appear without a clear first conjunct, in which case we can assume the first conjunct to be a salient proposition (explicit or not) in previous discourse. Unlike adverbials, these introductory coordinators are placed before the first position in the V2-sentence (22).

(22) Det är därför han dabbar sig ibland. För han vågar chansa. (a)
because he dares take risk

'That's why he messes up sometimes. Because he's not afraid to take a risk.'

When the conjuncts are clausal, issues in distinguishing coordination (conjuncts plus a coordinator) from subordination (dominating material, subordinated clausal material, and a subordinator) arise. We then need to refer to several additional properties of coordinators and subordinators. First, certain coordinators allow more than two conjuncts, whereas subordinators never relate more than two syntactic units. Secondly, depending on the subordinator, subordinated material may be placed at the front of a clause, which means the subordinator linearly precedes both the subordinated and the dominating material — coordinators on the other hand must appear between conjuncts.[14] Finally, subordinated material may always (sometimes: must) be realized with subordinated clause word order, whereas in coordinations, there is parallelism between the conjuncts, or, for certain coordinators, the conjunct following the coordinator is constrained to main clause order.

4.9 Numerals

We gather non-inflecting cardinal numbers into a separate part of speech, numerals, for which there are no specific features. Numerals are typically suitable for denoting specific quantities. Distributionally, they behave like other quantity denoting words, such as *få* 'few', *många* 'many' (see SAG 6 §1). SAG calls these latter two 'quantity pronouns', whereas we consider them to be adjectives. Numerals, however, distinguish themselves from pronouns and adjectives by their lack of inflection and by their systematic exocentric compounding behaviour. Not all numerals denote specific quantities (*femtioelva* 'umpteen', lit. 'fifty-eleven'), nor do all uses of numerals necessarily express specific quantities (*femtio pers* 'fifty persons', could be used to describe an approximate number of people).

Not included in the class of numerals is the word for 'one', *en/ett*, which we treat as a pronoun. We refer to Section 5.1 for a more detailed discussion of this decision. Certain other words with the potential of denoting specific quantities like *dussin* 'dozen', *gross* 'gross' are common nouns, and not numerals. Likewise, expressions for high powers of ten such as *miljon* 'million', *miljard* 'billion' are nouns, because of their inflection and distribution characteristics (SAG 6 §6).

We consider ordinal numbers to be adjectives (like UDP, but unlike SAG, SUC, and MAMBA). This means that there is a systematic relation between numerals and adjectives: *tjugotre* 'twenty three' (numeral), *tjugotredje* 'twenty third' (adjective). There are also systematic links between numerals and nouns, through the suffix *-a*: *tjugotre-a* 'number twenty three' (noun); and (via the ordinals) by compounding with *del* 'part': *en tjugotredjedel* 'one twenty-third' (noun, lit. 'a twenty-third part').

[14] That is, we find the linear order *sub X Y*, with structure [[*sub X*] *Y*], whereas the linear order *coord X Y* does not occur. See also Haspelmath (2007) for a typological generalization and discussion of the contrast between coordination and subordination. The case of discontinuous coordinators is not a counterexample to this generalization. Although, for instance, *varken du eller jag* 'neither you nor me' appears to be of the form *coord X coord Y*, we cannot treat the initial part *varken* alone as the coordinator: **varken du jag*. We argue that the proper annotation of a discontinuous coordinator involves analysis as a multiword unit. The latter are discussed in Section 6.

Ordinals and cardinals written with digits are always treated as numerals and adjectives, respectively: *5 kr* '5 SEK' (numeral), *den 5 januari* 'the 5th of January' (adjective). For the sake of simplicity, this rule also applies to *1* (here not a pronoun), *144* (here not the noun *gross*), and *1 000 000* (here not a noun phrase consisting of the pronoun *en* and the noun *million*).

4.10 Interjections

Interjections are non-inflecting words that are characterized by their lack of integration with their syntactic surroundings. For example, they do not trigger subject-verb inversion, unlike adverbs of similar content. Instead, interjections typically have the distribution of an independent utterance. Many interjections are used to emphasize expressive rather than descriptive content (*ack* 'o!', *bu* 'boo', *usch* 'yuck'), others may be used for social conventions (*hejdå* 'goodbye'), as discourse contributions or regulators (*ja* 'yes', *nej* 'no', *hursåå?* 'beg your pardon?', also filled pauses), or as onomatopoeia (*pang* 'bang', *mu* 'moo', *klick* 'click'). We follow SAG (12 §15) in accepting that some interjections take optional complements, as in *Hej på dig!* 'Hello, you!'. We do not define any specific features for interjections.

There are some systematic exceptions to the non-integration generalization. In direct speech, the reported material functions as a constituent in the reporting clause (23a). Onomatopoeia may be used with verbs like *låta* 'sound' (23b), and emotive interjections can be used predicatively (23c).

(23) (a) Hej, sa Petronella ifrån Plaskeby[.] (a)
hi said Petronella from Plaskeby
'Hi! said Petronella from Plaskeby.'

(b) Brum brum, låter bilen. (a)
vroom vroom sounds the car
'Vroom vroom goes the car.'

(c) Känner mig blä just nu[.] (a)
feel REFL yuck right now
'I feel yuck (= don't feel well) right now.'

We consider all of these to be interjections, nevertheless. On the other hand, secondary interjections (Ameka, 1992) — verbs such as *Akta!* 'Beware!', adverbs such as *Tyvärr!* 'Alas!/Sadly so!' — are not annotated as interjections, unless there are clear differences between them in pronunciation, distribution, or meaning, in addition to their use as independent utterances.

4.11 Symbols

Table 8: Symbol-specific features

Type	delimiter, other symbol

The category of symbols gathers non-alphabetic characters and character combinations, mainly punctuation marks, and emoticons and emojis. Although these are not generally treated in a grammatical description, since they partially fall outside of the realm of the morphosyntactic, a part-of-speech inventory that is intended to annotate running text has to provide a way to deal with these. Having a way of handling such items benefits both further (automatic) processing, and subsequently the usability of a corpus.[15]

As shown in Table 8, we use one symbol-specific feature. We distinguish punctuation — such as commas, semi-colons, and parentheses — from other symbols by annotating the type feature as delimiter. All other symbols are annotated with the (uninformative) label 'other symbol'.

Pictographs (emojis, smileys, etc.) are sometimes used in a syntactically non-integrated fashion — in that sense they resemble interjections — but they can also be used as more or less regular syntactic items. It would then seem to make sense to analyze them after their use: The underlined heart symbol in *I <3 creepy crawlies* would be verb, whereas the same in *all my <3 to you* would be a noun. However, far too often, this reasoning fails to give a clear, unambiguous result. Note that we have no information from inflection, and a distributional substitution test may have many answers, depending on how we choose to 'read out' the pictograph. Consider the attested example in (24), which would be compatible with an analysis of the two emoticons as nouns ('happiness'/'sadness'), adjectives ('happy'/'sad'), verbs ('smile.INF'/'frown.INF'), or interjections ('yippee'/'boohoo').

(24)	Hur	man	går	från	:)	till	:(	på två sekunder		(a)
	how	one	goes	from	(OTHERSYM)	to	(OTHERSYM)	in two seconds		

'How to go from :) to :(in two seconds.'

Abbreviational symbols are often more broadly conventionalized and frequently figure in an integrated fashion. Examples are *%* for *procent* 'percent', *&* for *och* 'and', and the dash – to indicate a range, read as *till* 'to'. We label these as symbols, but we also note that in our related syntactic annotation schema, these symbols are allowed to participate in the syntactic trees in the appropriate way.

Whenever possible, tokens that are combinations of symbols and morphemes (according to a recognizable orthography) are not treated as symbols but as the appropriate other part of speech: *tack <3:at* 'thank you, dear' (lit.: heart symbol for Swedish *hjärta*, linking grapheme ⟨:⟩, and *-at*; to form *hjärtat* 'heart(N).DEF.SG')

[15]The presence of symbols, and the need to deal with them in some way, is prominent in social media material, which is also present in our development corpus.

SAG does not discuss symbols since these are not traditionally considered to be a part of grammar. See, however, Nunberg (1990) for a different perspective. Neither SUC nor MAMBA have a category for symbols, although they both have categories for punctuation. SUC further distinguishes different types of punctuation, which we do not. UDP has two different categories for punctuation and symbols. Both SUC and UDP tag abbreviation symbols with a regular part of speech. For instance, *$* would be a noun, following *dollar*.

4.12 Foreign Material

The class of foreign language material exists for the same reasons as the symbol class: In an annotation setting, we are confronted with such items and need a way to deal with them. Our conception of foreign material is broad, it not only covers interspersed foreign natural language, but also formal language such as programming languages, in-line musical notation, formulae from mathematics and logic, and IPA notation. There are no specific features associated with foreign material. The category is also used in SUC, while UDP only has the category 'other'.

Where it is possible to give a more specific part of speech, for instance because a lexical item of foreign origin can be considered as part of the Swedish vocabulary, or when it is clear that we are dealing with a proper name, those labels are preferred over the foreign word label. Furthermore, visibly inflected material is, analogous to the case of symbols, treated as belonging to one of the regular word classes. Consider the example in (25).

(25) Om det fanns Nobelpris i att carp-a diem [...]. (a)
if it existed Nobel prize in to carp-INF diem

'If there was a Nobel prize for seizing the day [...].'

Here, the use of the infinitival suffix *-a* in *carpa* means it gets annotated as a verb, whereas *diem* receives the label foreign word.

5 A Few Special Mentions

We have now described the parts of speech of the Koala tagset. There are, however, a few further issues that require elaboration, since our treatment of these differ from other descriptions of Swedish. This section reviews two words, our pronoun treatment of the numeral sense of *en* and the relativizer *som*, and one category, SUC's verb particles.

5.1 The Numeral *one*

The word *en* and its inflections show up in several uses or senses, possibly associated with different lexical items. Here we wish to discuss two of them in particular: the indefinite article, and the numeral sense 'one'. Both inflect in the singular indefinite as *en* SG.IND.C, *ett* SG.IND.N. In the part-of-speech inventory presented here, we consider both senses to be pronouns. This can be contrasted with SUC, which categorizes *en* either as a special cardinal with gender distinction, as a determiner, or as a pronoun; and with UDP; which also associates the two senses of *en* with different parts of speech, namely numeral and

determiner. MAMBA gives *en* a category of is own, thus recognizing that the two senses are special or difficult to distinguish.

SAG discusses *en* both as one of the cardinal numerals (SAG 6 §2ff) and as an indefinite article (SAG 5 §166ff). At the same time, in both locations, they are stated to be one and the same, and SAG does not offer any means of distinguishing the two.[16] This is a clear example where SAG does not need to clearly define the borders between categories, since it is not meant as an annotation guideline.

As mentioned, we consider both the numeral sense *en* and indefinite article *en* to be pronouns. The motivation for this is the fact that *en/ett* inflects for gender, whereas none of the members of our part-of-speech category of numerals does. We consider a word's inflection properties to be our primary categorization criterion (see the discussion in Section 3). The alternation *en–ett* clearly resembles that of adjectivally inflected pronouns in form and distribution. Interestingly, compound numerals ending in *-en* or *-ett* do not show the same type of agreement. The *-en* form is the generally applicable form that can be used with common gender as well as neuter heads (26a). The *-ett* form is used in a restricted set of contexts, for instance when the head is *år* 'year', or when specifying *nummer* 'number' (26b).

(26) (a) sextioen hus / villor
sixty-one house(N).PL.IND house(C).PL.IND
'sixty one houses'

(b) Låten nådde nummer sextioett på Bilboard Hot 100 [.] (a)
song.DEF reached number sixty-one on Bilboard Hot 100
'The song reached number 61 on the Bilboard Hot 100 chart.'

Compound numerals with *-en* can basically be considered to be non-inflecting,[17] which means they fit in as numerals.

A possible objection against labeling numeral-sense *en* as a pronoun is that it can be coordinated with a numeral, as in: *en eller två* 'one or two'. However, coordination is not a good test for part-of-speech membership, since conjuncts may be of different categories (see also the remarks in Section 4.8). So, in addition to *en eller två*, we have, for instance, *en eller fler* 'one or more', and *en eller några* 'one or several'.

A related but more general argument is that our treatment of *en* goes against the intuition that a word with a distribution like the numerals, that expresses a specific cardinality should, indeed, be a numeral. However, the differences in distribution between numerals and pronouns are not enough to decide between the two categories. Ultimately, this argument is then based on semantics. We argue that placing *en* in a different part of speech than the words for 2, 61, or 100 is fully compatible with acknowledging the existence of a clear numeral sense for it. We simply do not consider this existence to be enough to overrule inflection as an organizational principle. After all, countable quantity expressions are not only found in our class of numerals, but also among pronouns: *några* 'several'; among adjectives: *få* 'few', *många* 'many'; and in the form of phrases headed by

[16] 'Den obestämda artikeln är identisk med grundtalet för 1.' (SAG II, 406), 'Den obestämda artikeln (som också är grundtalet för 1) [...].' (SAG II, 408), and 'Grundtalet *en* (= den obestämda artikeln) [...].' (SAG II, 480)

[17] See also `https://www.sprakochfolkminnen.se/sprak/sprakradgivning/frageladan.html` under the question *När använder man tjugoen och när tjugoett?*, consulted 1 September 2019.

nouns: *ett tiotal* 'roughly ten', *en miljon* 'a/one million'. These are all examples — in our system — of non-numerals, with senses and distributions that overlap with those of the numeral-sense *en*.

5.2 The Relativizer *som*

The word *som* has an interesting category history, which can be understood if we look at its distribution. It shows up in a variety of contexts, such as comparison (cf. English *as* and *like*), relative clauses (cf. English relativizer *that*), subordinate interrogative clauses, conjunction (cf. English *(as well) as*), in predicatives, and in certain cases as a governed element (*fungera som* 'function as', *betrakta som* 'regard as/to be').

The main difference between the discussed guidelines is whether *som* used in relative and interrogative contexts is considered to be a relative/interrogative pronoun taking an argument function, or whether it is seen as a functional element or marker. SUC, UDP,[18] and MAMBA take the former route, and accordingly tag *som* in these cases as interrogative/relative pronoun. When an adverbial is relativized, SUC marks *som* as interrogative/relative adverb.

For the remaining uses of *som*, we find a mix of strategies. SUC marks *som* as a subjunction when followed by clausal material, and puts it in the same group as the (coordinating) conjunctions when followed by nominal material. UDP marks the other *som*-uses as preposition, subordinating conjunction or coordinating conjunction. MAMBA labels *som* as subordinating or coordinating conjunction in these cases.

We follow SAG in considering most of the mentioned uses of *som* to be subordinators. See Stroh-Wollin (2002) for arguments why *som* in relative clauses is not a pronoun. We do not distinguish between a preposition *som* (followed by a noun phrase or other non-clausal material) and a subordinator *som* (followed by a clause/clausal fragment), since in many cases it would involve difficult reasoning during annotation about the extent to which the complement can be understood to be the result of comparative deletion.

All guidelines and descriptions considered here, including ours, agree upon analyzing *som* in conjunctions as a coordinator: *barn som vuxna* 'children and adults alike'.

5.3 Verb Particles

Particle verbs are common in Swedish, and the SUC guidelines assume a specific part of speech for verb particles, which sets it apart from the other descriptions. Verb-particle combinations share one word accent, which falls on the particle, and may show non-compositional semantics (see also Section 6 on multiword units). In many cases, the particle is identical in form with an adverb (*slå igen* 'close', lit. 'hit against'), or a preposition (*hoppa i* 'jump in'). But some are combinations with nouns (*äga rum* 'happen/take place', lit. 'own space'), adjectives (*sitta fast* 'be stuck', lit. 'sit fixed') or verbs (*låta bli* 'omit/leave be', lit. 'let become'). Verb particles notably differ from prepositions in their placement with respect to any further complements, as well as their ability to appear without their own complement, even when they are identical in form with prepositions.

[18] Our statements here about UDP's treatment of *som* are inferences from the available treebank annotations.

Unlike their counterparts in other functions, some particles have adverbial-like inflection, see the contrast in (27), where sentence (a) has particle placement of the adjective, and sentence (b) has secondary predicate placement. Recall that the indefinite singular neuter inflection for adjectives is standard with adverbial use (Section 4.4).

(27) (a) göra färdig-t / *färdig-a artiklarna (s)
make done-SG.IND.N done-PL.IND|DEF.C|N article.PL.DEF.C
'finish the articles'

(b) göra artiklarna färdiga (s)

The inflectional, distributional and semantic idiosyncrasies of verb particles can only be understood by recognizing that they are verb particles. From that perspective, SUC's choice to put these in a separate class makes sense. However, we follow SAG in recognizing a grammatical function of particle adverbial at the syntactic level (also see Josefsson, 2005, p. 47, for a discussion of both sides). For us, this removes the necessity, and thereby the benefit, of modeling these facts at the level of part of speech. In our view, occurrence as a verb particle cannot be used as evidence for membership of any of the parts of speech. UDP and MAMBA treat verb particles in a similar manner to ours.

A remaining problem is the status of verb particles that only occur as such, and nowhere else, for instance *komma ihåg* 'remember/remind oneself' (lit. 'come *ihåg*') or *slå dank* 'dawdle' (lit. 'hit *dank*').[19] In these cases, we use ad hoc (and largely immaterial) part-of-speech assignments, based on circumstantial reasoning. For instance, *ihåg* is an adverb since it looks like a fused prepositional phrase with *i* 'in' and *håg* 'mood'; and *dank* is a noun because of the homonymous noun *dank* 'steel marble', even though the relation is unclear. These cases are easy to list and such decisions therefore only have to be made once.

6 Multiword Units

The discussion so far has focused on annotation units that can be said to be single syntactic words — even though these may consist of multiple graphic words. In this section, we briefly discuss multiword units, that is, combinations of words that display idiosyncrasies with respect to their individual parts. Multiword units are relevant here because of our assumption that we can categorize these according to the same model as parts of speech, that is, as lexical items, using the same label inventory for them as for single word units (cf. Borin et al., 2013).

We have already discussed the idiosyncratic distribution of meta-linguistic material (Section 4.2). This may show up as head of a noun phrase, irrespective of its "original" part of speech. We solve this by analyzing such cases as common nouns. Metalinguistic material may consist of multiple words, sometimes even whole sentences, and gathering them under a multiword common noun label captures their noun-like behaviour. A similar argument can be made for titles behaving like proper names.

A change of category not only happens with meta-linguistic material and titles. It is particularly common with prepositional phrase multiword units. So, *bland annat* 'amongst

[19]These are typically easy to analyze from a historical perspective, but can be more or less opaque to contemporary speakers.

other (things)' and *på momangen* 'now/immediately' (lit. 'at moment.DEF') are arguably multiword units because of their collocational status and as a unit they are adverb-like. On the other hand, *från vettet* 'crazy' (lit. 'away from sense.DEF'), *på örat* 'drunk' (lit. 'on ear.DEF') behave like predicative adjectives and are thus considered adjectival.

Multiword units need not be contiguous, and can be more or less syntactically flexible (Sag et al., 2002). For instance, Swedish has a small number of circumpositions, such as *för ... sedan* 'since', *åt ... till* 'towards', which we analyze as multiword adpositions. Likewise, discontinuous coordinators are taken to be multiword coordinators (*såväl ... som* 'as well ... as'). We also analyze particle verbs (*komma ihåg* 'remember'), obligatorily reflexive verbs (*lata sig* 'be lazy', lit. 'laze oneself') and (semi-)fixed verb-complement combinations (*vidta åtgärder* 'take measures') as multiword verbs. The verb and its complement may be separated, and in some cases the complement can be inflected or syntactically modified.

The balance between the different types of arguments for a multiword unit's category differs from the single word case. In particular, we rely much more on distributive and even semantic information when we wish to establish the part of speech of a multiword unit. This is because inflection may be highly atypical: Prepositional phrase multiword adjectives do not inflect at all. A multiword noun of the form adjective+noun may have multiple exponence of agreement inflection, with adjective-like inflection on the adjective, and noun-like inflection on the noun: *röd blodkropp* 'red bloodcell', *röd-a blodkropp-ar* 'red-PL bloodcell-PL'. Some multiword verbs inflect in the dimensions of the verb as well as in the dimensions of the complement: *vidtar/vidtog åtgärder/-na* 'takes/took measures/measures-DEF'. In terms of inflectional paradigms, the result is a product of the verb's paradigm and the complement's paradigm. What is more, when the complement is a reflexive pronoun, there has to be agreement between the pronoun and a subject. Seen as a unit, then, the multiword verb that contains a reflexive shows a form of subject-verb agreement: *jag latar mig*, *du latar dig*, 'I am lazy', 'you are lazy', lit. 'I laze myself', 'you laze yourself', etc. This is in spite of the fact that, otherwise, contemporary Swedish verbs do not show agreement with their subjects.[20]

In our combined part-of-speech and syntactic annotation, multiword units receive double annotation: We annotate the parts as single words and the whole as a node in the *syntactic* structure, albeit with a *lexical* label. The double analysis means that the occurring inflection is annotated on the parts, while we do not mark any features on the multiword node. This has the (technical) advantage that we do not have to introduce new paradigms for multiword units in cases like the multiword verbs mentioned in the previous paragraph. The node, however, does give us a point where we can attach the multiword lemma. It also functions as a flag for the distributional changes that recognition as a multiword unit entails. We refer to Adesam et al. (2015b) for a longer discussion of our treatment of multiword units in the syntactic analysis.

[20]We can illustrate the balance shift like this: When confronted with a unit like *vidta åtgärder* 'take measures', inflection is not a clear argument for its categorial status, since verbs do not inflect like *åtgärder, åtgärderna*, and nouns do not inflect like *vidta, vidtar, vidtog, vidtagit* etc. However, looking at distribution, we notice that we can form a root sentence by combining a pronoun in subject form, like *jag*, with the (inflected) form *vidtar åtgärder* — which is characteristic of verbs, and not of nouns. Therefore, *vidta åtgärder* is a multiword verb.

7 Conclusions and Outlook

This article describes and motivates the Koala part-of-speech tagset. We distinguish 13 coarse-grained categories, which are further refined by a set of features marking various types of inflection, morphological properties, or sub-categories. The tagset is an integral part of a larger annotation schema for the Swedish language, which also covers syntactic structure and lexical semantics. These layers are closely connected. We have annotated a development corpus with this schema, which covers material from different contemporary sources and genres. The part-of-speech tagset is based on existing grammatical descriptions, mainly the Swedish Academy Grammar. One of our main goals has been to arrive at an annotation model which fits into the field and tradition of modern descriptive Swedish grammar.

In the design of our tagset, we have focused on morphologically visible distinctions. As a consequence, a word's inflectional properties are the main criterion in determining its part of speech. This focus has also influenced the type and number of morphological features we include in the description. The result is a fairly compact set of tags and features. This may not always be enough for researchers with different interests, and there may be a need for a more detailed description. We would like to think that our current proposal could serve as a good starting point, a base to which such further details can be added. Moreover, we envision adaptations of the tagset for other types of Swedish. For example, we would like to explore using our annotation schema for spoken language in the future.

In the article, we have compared our part-of-speech categories to other available descriptions. In ongoing work, we concentrate on creating, potentially lossy, mappings between the Koala tagset and other descriptions, starting with SUC and UDP.

One issue which we have barely touched upon is the automatic application of the tagset. Several aspects of the annotation, such as allowing for spaces in tokens, present challenges for computational processing. To us, however, the computational issues are secondary compared to the linguistic relevance. Only with such a description model in place can we start working on its automatic applications.

Acknowledgements

The Koala project was funded 2014–2017 by Riksbankens Jubilemsfond, grant number In13-0320:1. This article would not have been written without fruitful discussions with a number of people. We would like to thank our project colleagues Lars Borin, Markus Forsberg and Richard Johansson, and are also very grateful for discussions with Elisabet Engdahl, Benjamin Lyngfelt, and Joakim Nivre. We also thank the three anonymous reviewers for their comments and questions. Finally, we are indebted to our annotators, who, besides annotating 100 000 tokens of text with moving target annotation guidelines, have pointed out discrepancies and many unforeseen uses of language.

References

Adesam, Yvonne, Gerlof Bouma, and Richard Johansson. 2015a. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the Nordic Conference on Computational Linguistics (Nodalida)*. Linköping University Electronic Press, Sweden.

Adesam, Yvonne, Gerlof Bouma, and Richard Johansson. 2015b. Multiwords, word senses and multiword senses in the Eukalyptus treebank of written Swedish. In *Proceedings of TLT*. ISBN 978-83-63159-18-4.

Adesam, Yvonne, Gerlof Bouma, Richard Johansson, Lars Borin, and Markus Forsberg. 2018. The Eukalyptus treebank of written Swedish. In Swedish Language Technology Conference (SLTC). Available at `https://sltc2018.su.se/program/`. Stockholm University.

Ameka, Felix. 1992. Interjections: The universal yet neglected part of speech. *Journal of Pragmatics* 18(2–3):101–118.

Baker, Mark and William Croft. 2017. Lexical categories: Legacy, lacuna, and opportunity for functionalists and formalists. *Annual Review of Linguistics* 3(1):179–197.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In Swedish Language Technology Conference (SLTC). Available at `https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf`. Umeå University.

Borin, Lars, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation* 47(4):1191–1211.

Borin, Lars, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, page 474–478. European Language Resources Association (ELRA).

Börjars, Kersti. 2003. Morphological status and (de)grammaticalisation: the Swedish possessive. *Nordic Journal of Linguistics* 26(2):133–163.

Carlberger, Johan and Viggo Kann. 1999. Implementing an efficient part-of-speech tagger. *Software: Practice and Experience* 29(9):815–832.

Croft, William, Dawn Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, pages 63–75.

Davis, Randall, Howard Shrobe, and Peter Szolovits. 1993. What is a knowledge representation? *AI Magazine* 14(1):17–33.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Durkin, Philip. 2015. *The Oxford Handbook of Lexicography*. Oxford University Press, 1st edn. ISBN 9780199691630.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project - description and guidelines. Tech. Rep. 33, Department of Linguistics, Umeå University.

Forsbom, Eva. 2008. Good tag hunting: Tagability of Granska tags. In J. Nivre, M. Dahllöf, and B. Megyesi, eds., *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, no. 7 in Studia Linguistica Upsaliensia, pages 77–85. Acta Universitatis Upsaliensis. ISBN 978-91-554-7226-9.

Fuertes-Olivera, Pedro A. 2017. *The Routledge Handbook of Lexicography*. Milton: Routledge. ISBN 9781138941601.

Haspelmath, Martin. 2007. Coordination. In T. Shopen, ed., *Language Typology and Syntactic Description*, vol. 2, pages 1–51. Cambridge University Press, 2nd edn.

Haspelmath, Martin. 2012. How to compare major word-classes across the world's languages. In *Theories of everything: in honor of Edward Keenan*, no. 17 in UCLA Working Papers in Linguistics, pages 109–130. Los Angeles: UCLA.

Haspelmath, Martin. 2015. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In J. Błaszczak, D. Klimek-Jankowska, and K. Migdalski, eds., *How categorical are categories? New approaches to the old questions of noun, verb, and adjective*, pages 273–304. Berlin, Boston: De Gruyter Mouton. ISBN 9781614514510.

Jackson, Howard, ed. 2013. *The Bloomsbury Companion to Lexicography*. Bloomsbury Companions. London, England: Bloomsbury. ISBN 9781441145970.

Josefsson, Gunlög. 2005. *Ord*. Lund: Studentlitteratur. ISBN 9144037260.

Knutsson, Ola, Johnny Bigert, and Viggo Kann. 2003. A robust shallow parser for Swedish. In *Proceedings of the Nordic Conference on Computational Linguistics (Nodalida)*.

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

Nivre, Joakim. 2014. Universal Dependencies for Swedish. In Swedish Language Technology Conference (SLTC). Available at `https://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_7.pdf`. Uppsala University.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA).

Nivre, Joakim, Beáta Megyesi, Sofia Gustafson-Capková, Filip Salomonsson, and Bengt Dahlqvist. 2008. Cultivating a Swedish treebank. In *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*. Uppsala University, Department of Linguistics and Philology.

Nivre, Joakim, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1392–1395. European Language Resources Association (ELRA).

Nunberg, Geoffrey. 1990. *The linguistics of punctuation*. No. 18 in CSLI Lecture Notes. Stanford: Center for the Study of Language and Information (CSLI). ISBN 0937073474.

Osborne, Timothy and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics* 4(1):17.

Pullum, Geoffrey. 2009. Lexical categorization in English dictionaries and traditional grammars. *Zeitschrift für Anglistik und Amerikanistik* 57(3):255–273.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In A. Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Stroh-Wollin, Ulla. 2002. *Som-satser med och utan* som *[Som-clauses with and without* som*]*. Ph.D. thesis, Uppsala University.

Svensén, Bo. 2004. *Handbok i lexikografi: ordböcker och ordboksarbete i teori och praktik. Andra, omarbetade och utökade upplagan*. Stockholm: Norstedts akademiska förlag. ISBN 9172272694.

Teleman, Ulf. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund: Studentlitteratur. ISBN 91-44-10721-8.

Teleman, Ulf, Staffan Hellberg, and Erik Andersson. 1999. *Svenska Akademiens Grammatik*. Stockholm: Svenska Akademien. ISBN 9172271264.

Trask, R. L. 1999. Parts of speech. In E. K. Brown and J. E. Miller, eds., *Concise encyclopedia of grammatical categories*, pages 278–284. Oxford, Amsterdam: Pergamon; Elsevier. ISBN 008043164X.

Vogel, Petra and Bernard Comrie, eds. 2000. *Approaches to the Typology of Word Classes*. Berlin-New York: Mouton de Gruyter. ISBN 3-11-016102-8.

Volk, Martin, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. SMULTRON (version 3.0) - the Stockholm multilingual parallel treebank. An English-French-German-Spanish-Swedish parallel treebank with sub-sentential alignments, `https://www.cl.uzh.ch/en/texttechnologies/research/corpus-linguistics/paralleltreebanks/smultron.html`, (consulted 22 October 2019).

Appendix: Overview of the Tagset

Here we list the Koala part-of-speech labels, as used in the Eukalyptus treebank, with additional distinguishing features. The labels may differ from those used in the article, since the labels in the article use English abbreviations. Tables with a horizontal divider have label descriptions in the top and allowed label combinations with examples below. ∗ marks attested combinations in the Eukalyptus treebank.

Table A1: Koala feature labels for any part of speech.

	FKN.-.-	abbreviation
	-.GEN.-	genitive
	-.-.ESM	partial ellipsis
∗	-.-.-	
∗	FKN.-.-	ex.
∗	-.GEN.-	husets
∗	-.-.ESM	höst-
∗	FKN.GEN.-	EU:s
∗	FKN.-.ESM	FN- [och NATO-medlemsskap]
	-.GEN.ESM	[lägenhetssäljaren och] -köparens
	FKN.GEN.ESM	GU- [eller Chalmersstyrelsens sida]

Table A2: Koala feature labels for adverbs, AB.

	POS.-	positive degree
	KOM.-	comparative degree
	SUV.-	superlative degree
	-.FRL	wh- or relative
∗	-.-	ideligen
∗	POS.-	länge
∗	KOM.-	snarare
∗	SUV.-	oftast
∗	-.FRL	när

Table A3: Koala feature labels for adjectives, AJ.

	POS/KOM/SUV.-.-.-.-	positive/comparative/superlative degree
	-.SIN/PLU.-.-.-	singular/plural number
	-.-.IND/DEF.-.-	indefinite/definite definiteness
	-.-.-.NEU/UTR/MAS.-	neuter/common/masculine gender
	-.-.-.-.FRL	wh- or relative
∗	POS.SIN.IND.UTR.-	kvinnlig
∗	POS.SIN.IND.NEU.-	livligt
	POS.SIN.IND.UTR.FRL	hurdan
	POS.SIN.IND.NEU.FRL	hurdant
∗	POS.SIN.DEF.MAS.-	berömde
∗	POS.SIN.DEF.UTR|NEU.-	högtidliga
∗	POS.PLU.IND|DEF.UTR|NEU.-	politiska
	POS.PLU.IND|DEF.UTR|NEU.FRL	hurdana
∗	POS.SIN|PLU.IND|DEF.UTR|NEU.-	tredje
∗	KOM.SIN|PLU.IND|DEF.UTR|NEU.-	större
∗	SUV.SIN.DEF.MAS.-	förste
∗	SUV.SIN|PLU.IND.UTR|NEU.-	minst
∗	SUV.SIN|PLU.DEF.UTR|NEU.-	bästa

Table A4: Koala feature labels for nouns, NN.

	NEU/UTR.-.-	neuter/common gender
	-.SIN/PLU.-	singular/plural number
	-.-.IND/DEF	indefinite/definite definiteness
∗	UTR.SIN.IND	rösträtt
∗	UTR.SIN.DEF	mandatperioden
∗	UTR.PLU.IND	släktingar
∗	UTR.PLU.DEF	hästarna
∗	NEU.SIN.IND	ljud
∗	NEU.SIN.DEF	arbetet
∗	NEU.PLU.IND	partier
∗	NEU.PLU.DEF	barnen
∗	UTR|NEU.SIN|PLU.IND|DEF	fjol (ad hoc)

Table A5: Koala feature labels for pronouns, PO.

	IND/DEF.-.-.-.-	indefinite/definite definiteness
	-.SIN/PLU.-.-.-	singular/plural number
	-.-.NEU/UTR/MAS.-.-	neuter/common/masculine gender
	-.-.-.SUB/OBJ/PSS.-	subject/object/possessive form
	-.-.-.-.FRL	wh- or relative
∗	IND.SIN.UTR.-.-	någon
∗	IND.SIN.UTR.-.FRL	vilken
∗	IND.SIN.NEU.-.-	inget
∗	IND.SIN.NEU.-.FRL	vilket
∗	IND.PLU.UTR|NEU.-.-	några
∗	IND.PLU.UTR|NEU.-.FRL	vilka
∗	DEF.SIN.UTR.-.-	all
∗	DEF.SIN.UTR.SUB.-	hon
∗	DEF.SIN.UTR.OBJ.-	honom
∗	DEF.SIN.UTR.SUB|OBJ.-	den
∗	DEF.SIN.UTR.PSS.-	min
∗	DEF.SIN.NEU.SUB|OBJ.-	det
∗	DEF.SIN.NEU.PSS.-	sitt
∗	DEF.SIN.MAS.-.-	denne
∗	DEF.PLU.UTR|NEU.-.-	alla
∗	DEF.PLU.UTR|NEU.SUB.-	vi
∗	DEF.PLU.UTR|NEU.OBJ.-	oss
∗	DEF.PLU.UTR|NEU.SUB|OBJ.-	dom
∗	DEF.PLU.UTR|NEU.PSS.-	mina
∗	DEF.SIN|PLU.UTR|NEU.OBJ.-	sig
∗	DEF.SIN|PLU.UTR|NEU.PSS.-	deras

Table A6: Koala feature labels for symbols, SY.

∗	DEL	delimiter	?
∗	SYM	symbol	;-)

Table A7: Koala feature labels for verbs, VB.

	IND/KON/IMP/SPM/INF.-.-	indicative/subjunctive/imperative/supine/infinitive
	-.AKT/SFO.-	active voice/s-form
	-.-.PRS/PRT	present/past tense
∗	IND.AKT.PRS	varnar
∗	IND.AKT.PRT	vilade
∗	IND.SFO.PRS	visas
∗	IND.SFO.PRT	utvecklades
∗	KON.AKT.PRS	vare
∗	KON.AKT.PRT	finge
	KON.SFO.PRS	
	KON.SFO.PRT	funnes
∗	IMP.AKT.-	tillåt
	IMP.SFO.-	bits
∗	SPM.AKT.-	utforskat
∗	SPM.SFO.-	stängts
∗	INF.AKT.-	vandra
∗	INF.SFO.-	uppfattas

Table A8: Koala part-of-speech labels without further specific features, with examples.

Proper name	EN	∗	Gösta
Interjection	IJ	∗	Fy
Coordinator	KO	∗	och
Numeral	NU	∗	tolv
Preposition	PE	∗	på
Subordinator	SU	∗	eftersom
Foreign word	UO	∗	wizard