Defining the Eukalyptus forest – the Koala treebank of Swedish

Yvonne Adesam Gerlof Bouma Richard Johansson
Språkbanken
Department of Swedish
University of Gothenburg
{yvonne.adesam, gerlof.bouma, richard.johansson}@gu.se

Abstract

This paper details the design of the lexical and syntactic layers of a new annotated corpus of Swedish contemporary texts. In order to make the corpus adaptable into a variety of representations, the annotation is of a hybrid type with head-marked constituents and function-labeled edges, and with a rich annotation of non-local dependencies. The source material has been taken from public sources, to allow the resulting corpus to be made freely available.

1 Introduction

Corpora annotated with part-of-speech tags and syntactic structure are crucial for the development and evaluation of automatic tools for syntactic analysis, as well as for empirical research in syntax. For Swedish, annotated corpora have been available for quite a number of years. The venerable MAMBA treebank (Teleman, 1974) was created in the 1970s. It has formed the basis for a number of Swedish constituency and dependency treebanks such as Talbanken05 (Nivre et al., 2006), the more recent Swedish Treebank, and the Swedish part of the multilingual Universal Dependency Treebank (de Marneffe et al., 2014). The Stockholm–Umeå Corpus (SUC) (Ejerhed et al., 1992) with manually checked part-of-speech tags and base forms for roughly a million tokens, has been a de facto standard for Swedish part-of-speech tagging. The Swedish Treebank uses the SUC part-of-speech tags together with the automatically converted syntactic structures from MAMBA (Nivre et al., 2008).

In our project Koala, we develop new annotation tools to be used for the multi-billion token corpora of Korp, the corpus query infrastructure at Språkbanken. Part of our effort lies in evaluation of these annotation tools. For a number of reasons, the corpora mentioned and their annotation schemata are not suitable as our gold standard.

First, the texts in the corpora are quite dated, and do not reflect the text types available in Korp. Secondly, the MAMBA annotation would require several complex conversion heuristics to be used as a conventional constituency or dependency treebank. Due to technical limitations in the 1970s, attachment in MAMBA is underspecified in some cases, most notably in clause coordination, and its annotation does not have explicit phrase categories. On the other hand, its set of grammatical function categories is very fine-grained, and we consider some more semantic/pragmatic distinctions hard to apply. For the Swedish Treebank we further note that the part-of-speech tags and the syntactic categories were designed in separate projects, and there are several cases of redundancy, where grammatical function distinctions are also reflected in the set of part-of-speech tags.

In this paper, we describe the design of the syntactic layer, and to some extent the part-of-speech layer, of the new *Koala* multi-genre annotated Swedish corpus. In designing the annotation guidelines, we have aimed to address the above-mentioned shortcomings: First, the part-of-speech, phrase, and function categories have received clearly separated roles. Secondly, we use a syntactic annotation format that is less restrictive than MAMBA's. Thirdly, the annotation model has been designed with deterministic conversion into other formalisms in mind. Finally, the corpus consists of material from several genres. The texts have been collected from public-domain sources, so that the corpus can be made freely available. With the data release, we will also supply scripts for conversion to other standards.

2 The Koala corpus

The Koala corpus will consist of at least 100k tokens of modern Swedish text of various types, with about 20k tokens of each different text type.

- Novels: the first chapters from four novels
- Wikipedia: full articles from Swedish Wikipedia, 3k to 100 tokens per article
- Blogs: blog entries from the SIC corpus (Östling, 2013)
- Europarl: proceedings from the European parliament (Koehn, 2002)
- News/community information: we would have liked to add news text, but due to IPR restrictions, this is mainly community information (government information, health service information etc.)

Sentence segmentation and tokenization is based on orthographic words and sentences. This does not rule out the possibility of having syntactic tokens that span several graphic words, as the syntactic annotation readily allows multiword expressions (see Section 4.3: 'multiword expressions'). Graphic words containing several tokens, each with their own syntactic contribution – such as *serunte* for *ser (d)u (i)nte* 'don't you see', lit. 'see you not' – do however receive special treatment. The texts are manually annotated using an adapted version of the Synpathy tool.[1]

3 Lexical annotation

The part-of-speech tag set is a reduced version of the SUC tag set, with alterations to make it more consistent with the Swedish reference grammar SAG (Teleman et al., 1999). The labels are listed in Table 1. Nouns are marked for gender, number, and definiteness. Adjectives are marked for degree (POS/KOM/SUV), gender, number, and definiteness. Adverbs are marked for degree and whether they are relative or wh-pronouns (+FR). Verbs are marked for mood/finiteness, voice (where we, following SUC, distinguish between active and s-form, rather than active, passive, deponent, etc.), and in the case of indicative and subjunctive we also mark tense. Pronouns are marked for gender, number, definiteness, form (subject, object or possessive), and wh/relative. Proper nouns, numerals, interjections, subordinators, coordinators, prepositions, and foreign words are not further specified. Symbols are divided into punctuation and other.

Traditionally, the nominative-genitive case distinction is made for nominal parts-of-speech. However, in Swedish *-s* can either be the genitive suffix or it can be a phrase marking clitic, appearing on any NP-final word. In Koala we handle both these uses at the lexical level, using a single GEN feature that can appear on any part-of-speech. The example in (1) shows a GEN-marked preposition.

Part-of-speech		Features
AB Adverb	degree	POS KOM SUV
	wh/rel	+FR
AJ Adjective	degree	POS KOM SUV
	gender	UTR NEU MAS
	number	SIN PLU
	species	IND DEF
EN Proper noun		
IJ Interjection		
KO Coordinator		
NN Noun	gender	UTR NEU
	number	SIN PLU
	species	IND DEF
NU Numeral		
PE Preposition		
PO Pronoun	gender	UTR NEU MAS
	number	SIN PLU
	species	IND DEF
	form	SUB OBJ PSS
	wh/rel	+FR
SU Subordinator		
SY Symbol	type	DEL SYM
UO Foreign word		
VB Verb	mod/fin	IND KON IMP SUP INF
	voice	AKT SFO
	tense	PRS PRT

Table 1: The Koala Part-of-speech tag set, with morphological features.

(1) gå till den man ska svara pås gästbok
PE.GEN
go to them one shall reply to's guest book
'go to the guest book of the person you want to reply to'

In addition, parts-of-speech are marked with specific morphological labels when they are abbreviations, or when they are the incomplete part in an elliptical coordination (such as the first part in *lång- och kortfristiga lån* 'long and short term loans', or *1930- och 1940-talet* 'the 1930s and 1940s').

Compared to SUC, several categories are removed. Wh-adverbs are added to adverbs, participles and ordinal numbers to adjectives, and the infinitival marker to subordinators. Determiners, wh-determiners, possessive pronouns, wh-pronouns, and possessive wh-pronouns are added to pronouns. Particles are no longer a separate category, the majority being adverbs or prepositions. Punctuation is subsumed into the category of symbols.

In addition to the part-of-speech and morphological tags, we link words to the large-scale semantic

[1] http://www.mpi.nl/tools/synpathy.html

lexicon SALDO (Borin et al., 2013), which provides us with a lemma, the inflectional pattern and a sense distinction. We also follow SALDO in assuming that there is a multiword counterpart to each of the parts-of-speech. In the Koala syntax annotation schema, these multiword expressions reside between the lexical and the phrasal levels.

4 Syntactic annotation

4.1 Formalism

The syntactic structures in the Koala annotation schema follow the format introduced in Skut et al. (1997). It uses rooted trees, the 'primary graph', with additional, 'secondary', edges. All tokens part of the syntactic structure must occur as leaf nodes in the primary graph. Internal nodes in the primary graph represent phrases or (in our schema) multiword expressions. Unlike traditional phrase structure trees, linear order is not part of the encoding and phrases may be discontinuous. Word order variants therefore need not lead to different trees.

Edges, primary as well as secondary, carry grammatical function labels. Secondary edges are used for various kinds of sharing of syntactic material. With secondary edges included, syntactic structures can in principle be unrestricted directed graphs, however, in Koala we avoid cyclic structures.

Tokens are non-empty string segments, and the formalism does not allow for empty categories such as traces or null-pronouns. Discontinuous phrases and secondary edges together take care of most of the need for empty material.

The format has proven its suitability in several treebanks, including the German NEGRA (Brants et al., 1999), TIGER (Brants et al., 2004), and Tuba-D/Z (in restricted form) (Telljohann et al., 2012) treebanks, the Dutch CGN (spoken) (Hoekstra et al., 2001) and Lassy (written) (van Noord et al., 2013) corpora, the Swedish-German parts of the SMULTRON parallel treebank (Volk et al., 2010), and the Swedish Treebank (Nivre et al., 2006). It allows us to combine descriptive adequacy with ease of human annotation. It also allows us to convert the structures into dependency grammar or phrase structure grammar with as few heuristics as possible. The format ideally encodes the combined information found in analyses from either of these traditions.

4.2 Descriptive content

Our analysis of Swedish syntax is for important parts based on MAMBA and SAG. MAMBA contains a mix of elements from dependency grammar, topological field analysis and phrase structure grammar (see also Nivre (2002) for a brief description). The bulk of the dependency types Koala recognizes is taken from MAMBA, although Koala uses a much smaller set, especially in the adverbial and attributive modifier domain. Much of the grammatical argumentation is taken from SAG, as well as the set of phrase types. Of course, a reference grammar and an annotation model have very different goals: Whereas SAG can give a piecemeal description of different grammatical levels and domains and merely point out difficulties, ambiguities or non-discrete categorizations, the Koala schema needs to allow the annotator to assign a single complete tree to an annotation unit. On the other hand, Koala leaves much underspecified. Especially the rich semantic and pragmatic distinctions present in a comprehensive language description such as SAG's have been left out of Koala's system of functions and categories.

Phrasal categories, heads, and pseudoheads

Any of the part-of-speech categories of Section 3 may be used to construct a phrase with arguments and modifiers. The relation between a phrase and its head daughter (HD) is constrained by the following three properties:

Uniqueness There is at most one head in a phrase.
Lexicality The head daughter is a (multi)word.
Projection The phrase's category is determined by the head daughter's part-of-speech

In some cases, we wish to construct a phrase around a head-like element that violates one or more of these constraints. We then use the label pseudohead (PH). All allowed uses of PH are specified in the schema. Phrases are in principle allowed to be (pseudo-)headless, either just in terms of the primary graph or completely. The situations in which this may occur are specified (as much as possible) in the schema. An important motivation for the head constraints is ease of conversion to a dependency format[2] and increased possibilities for automatic error mining of the annotations.

[2]Of course, having a headed tree per se does not help in conversion to a format that uses different criteria for which part of a phrase functions as head.

Category		Head	Pseudohead
S	Sentence	VB.IND|KON|IMP	
VP	Verb phrase	VB.SUP|INF	
NP	Noun phrase	NN, PO, EN	AJ, AjP, NU, NuP
NuP	Numeral phrase	NU	
KoP	Coordinator phrase		KO
SuP	Subordinator phrase	SU	PO.+FR, AB.+FR, or AbP, NP, AjP, PP dominating such
PP	Preposition phrase	PE	
AjP	Adjective phrase	AJ	
AbP	Adverb phrase	AB	
IjP	Interjection phrase	IJ	
— any of the above —		— UO, SY.SYM —	

Table 2: Phrase categories and head projection rules. Note that wherever a part-of-speech is listed, its multiword counterpart is also accepted.

The inventory of phrase labels, and the part-of-speech tags they are projected from, are given in Table 2. The set of phrases largely follows SAG, although notably, unlike SAG, we do not recognize a finite VP, but instead combine the finite verb with its subject and other dependants directly in S.

We allow both function words (functional parts-of-speech) and content words (lexical parts-of-speech) as heads, unlike for instance the Universal Dependency Treebank (de Marneffe et al., 2014), which for reasons of cross-linguistic parallelism prefers content word heads. To illustrate, we distinguish a PP from an NP, instead of attaching the preposition as a case-like marker in the NP; and we recognize the level of SuP (subordinator phrase) rather than considering the subordinator to be a marker on one of the verbal projections. Although the majority of cases can straightforwardly be converted to a content word head-oriented annotation, we do note that in the case of a PP which embeds another PP or a SuP another SuP, we do not lose the hierarchical structure if we consider the PE or SU to be the head. Examples of the two annotation styles are in (2). The Koala annotation (2a) explicitly encodes the hierarchical information. The alternative – on the assumption of head lexicality – is the flat (2b), where the hierarchical information is only encoded in the linear order of the markers.

(2) a. [PP sedan HD [PP innan HD jul]]
PE PE NN
since before christmas
'since before christmas'

b. [NP sedan MARKER innan MARKER jul HD]]

In the same vein, we annotate modal and auxiliary verbs as heads rather than the main verbs, and copulas rather than the predicative complement (see also Section 4.3: 'the verbal domains').

The parts-of-speech SY and UO appear to violate the projection constraint: they may head any type of phrase and therefore do not determine the containing phrase's category. However, because of SY and UO's special status as marking lexical material outside Swedish morpho-syntactic conventions, they function as part-of-speech wild cards, and we do not consider phrases headed by SY or UO to violate projection. For instance, in (3), we have a symbol SY functioning as a verb heading an S, and a foreign multiword UOM functioning as a noun heading an NP.

(3) a. [S :'(HD inte för mig!]
SY AB PE PO
— not for me
'Don't cry for me!'

b. Det där är [NP ett [*sine qua non.*]HD]
PO UOM
that is a —
'That is a *conditio sine qua non*.'

We use pseudoheads PH for head-like daughters in three types of phrases: coordinators in coordinations (Section 4.3: 'coordination'), non-subordinator material introducing relative clauses or subordinate questions (Section 4.3: 'subordinate clauses') and adjectives or numerals in headless NPs (Section 4.3: 'the noun phrase').

Finally, unary branching nodes are avoided. So, bare nouns, adjective phrases, pronouns or numerals can serve directly as, say, direct object, without intermediate NP node. Likewise, we do not posit a unary SuP for bare subordinate clauses – they are simply marked S. In (4), an AjP (arguably with nominal flavour to it) directly serves as object (OO).

(4) Det gäller enbart [$_{AjP}$ nyligen anställda.$_{HD}$]$_{OO}$
AB AJ
this applies only newly employed
'This only applies to new employees.'

By exception, the primary graph will contain unary branching nodes when they are needed to accommodate secondary edges. See (12) in Section 4.3 for an example.

Edge labels

Koala uses a set of 2 head labels, 18 grammatical functions, and 2 extra-syntactic functions. All of them can appear in the primary or secondary graph. Some grammatical functions appear in different phrase types, possibly with subtly different meanings. For instance, we use a rather general MD label for modifiers in any domain. This contrasts with the tradition where such modifiers are called *attributive* in the nominal domain but *adverbial* in other domains. Similarly, the label OO is used for (direct) objects of verbs and adjectives (e.g., *likt dig* 'resembling you', lit. 'alike.NEU you'), as well as the complement (Swe: *rektion*) of subordinators or prepositions. Not all grammatical functions appear in every phrase type. Table 3 lists all functions and their domains. In the table, the designation '*' means any phrasal node can have an outgoing edge with the label in question, '*M' refers to the special multiword nodes, which are lexical pre-terminal nodes (see Section 4.3: 'multiword expressions').

Basic usage of some of the grammatical functions is illustrated in the sections below. The two extra-syntactic functions ME and DF fullfill rather different roles. They are used to include material in the primary graph that would otherwise not enter it for lack of syntactic interactions between the material and the rest of the graph. The multiword element label ME (see Section 4.3: 'multiword expressions') combines leaf nodes into a node to be used as one lexical unit. The discourse function DF is used for material that is coupled to the utterance, but does not relate syntactically to it: vocatives, left/right-dislocations, parentheticals, decorations, tags, etc. Example (5) shows the use of DF for the interjection *eller*, a homonym with coordinator 'or', which has an established use as a question tag.

(5) [$_{S}$ E$_{HD}$ du go, eller?$_{DF}$]
VB PO AJ IJ
are you good or
'Are you out of your mind!?'

Label	Meaning and domain
HD	head: *
PH	pseudohead: NP, SuP, KoP
SB	subject: S
	raised subject: VP, only secondary
ES	extraposed subject, pivot: S, VP
OO	direct object: S, VP, AjP
	complement: PP, SuP
EO	extraposed direct object: S, VP, AjP
	extraposed complement: PP
IO	indirect object: S, VP, AjP
AG	demoted subject in passive: S, VP, AjP
AN	bound apposition: NP
	free apposition: * not IjP
KL	conjunct: KoP
DT	determiner: NP
IV	non-finite verbal complement
	with raised subject: S, VP
JF	comparison: S, VP, AjP, AvP
MD	modifier: *
PL	verb particle: S, VP
OA	adverbial complement: S, VP, AjP, AvP
OP	object-oriented predicative: S, VP, PP
SP	subject-orentied predicative: S, VP
RA	locative/directional adjunct
	or complement: S, VP
EF	subordinate part of cleft: S, VP
ME	element of a multiword: *M
DF	discourse function: *

Table 3: List of edge labels, and their meaning per applicable domain

The discourse function can admittedly be (ab)used to make the graph span a unit defined in other terms – for instance the graphic sentence – by choosing a main part and adding the rest to it as DF-daughters. In any case, the parts that are connected by DFs are themselves syntactically coherent units, à la Loman and Jörgensen's (1971) *macrosyntagm*.

4.3 Selected applications

Below we give examples of how the system of phrases and functions is employed in some prominent domains of the grammatical system.

The verbal domains S and VP

The phrases S and VP project from respectively finite and non-finite verbs. Phrases of the category S typically contain at least a subject and a verb, further complements and adverbial modifiers of the verb can all be attached in the S node. We use a flat S annotation, irrespective of word order: V2 main clauses, (unmarked) SVO subordinate clauses, and V1 questions, imperatives or conditionals are all S.

VPs are built around a non-finite verb, and never contain their own (formal) subject. For subject con-

trol, we use a secondary SB edge, in which case the whole VP receives the special IV function. Arbitrary implicit subjects are not marked at all.

Example (6), a V1-imperative with an *accusativus cum infinitivo*, shows both an S node and a VP-node, and illustrates the use of a secondary edge for the VP's subject.[3]

(6) [S Snälla$_{DF}$ hjälp$_{HD}$ mig$^{\boxed{1}}_{OO}$ [VP $\boxed{1}_{SB}$ fatta.$_{HD}$]$_{IV}$]
IJ VB PO VB
please help me understand
'Please, help me understand.'

Because of the gradual nature of the auxiliary-main verb distinction in Swedish, we treat auxiliaries as embedding, just like any other verbal complement taking verbs. For instance, composite tense is treated as control using the IV function and a secondary subject in the embedded VP. Non-finite verbal material marked with *att* 'to' is annotated as a SuP containing a VP, with the infinitive marker heading the SuP.

The noun phrase NP

Noun phrases are projected from nouns, pronouns or proper names. The determiner role DT is specific to NPs, and is used for attributes of definiteness (including possessives) and quantity. Otherwise, the MD function is used as a general label for attributive material. In (7) we see a full NP, with both a definiteness and a quantity attribute, and with a prenominal adjectival modifier and a postnominal relative clause.

(7) [NP de$_{DT}$ två$_{DT}$ bästa$_{MD}$ låtar$_{HD}$ [S han gjort]$_{MD}$]
PO PO AJ NN PO VB
the two best songs he made
'his two best songs'

When an NP lacks a head in a coordination or more generally in ellipsis, we leave it without a head daughter in the primary graph completely, in coordinations the head is indicated using a secondary edge. Some NPs can be argued to construct around a non-nominal core, and annotating these as headless would be undesirable: realization of such NPs without a nominal head is the typical or even only way. Consider the AjP in example (4) above. The adjective *anställd* 'employed', without any nominal head, is the standard way of referring to an employee in Swedish. When combined with a determiner, as in (8a), we know we are dealing with an NP. We thus build an NP on basis of the AjP, and use the PH label to indicate that projection and lexicality are violated.

(8) a. [NP de$_{DT}$ [AjP nyligen$_{MD}$ anställda$_{HD}$]$_{PH}$]
PO AB AJ
the newly employed
b. [NP de$_{DT}$ nya$_{MD}$ anställda$_{PH}$]
PO AJ AJ
the new employed
'the new employees'

In (8b), we see a variant in which the NP with an adjective pseudohead contains an attributive premodifier.

Subordinate clauses S and SuP

Subordinate clauses fall in one of two categories, depending on whether they have pre-adjoined material marking them as subordinate clauses or whether they are bare. First, bare subordinate clauses are labeled S, as in (9).

(9) Jag tror [S jag$_{SB}$ är$_{HD}$ kär.$_{SP}$]
PO VB AJ
I think I am in love
'I think I'm in love.'

Embedded sentences may have a different word order than main ones, but, as mentioned, this does not change the categorization.

Secondly, embedded clauses are labeled SuP when they are introduced by a subordinator (10a) or by a wh- or relative-marked constituent (10b). Note that the latter is never an SU and may be phrasal. The two types of SuP-introducers are also distinguished by whether they have a syntactic function in the S embedded in the SuP. Note the secondary edge in (10b).

(10) a. Jag tror [SuP att$_{HD}$ [S du$_{SB}$ förstår.$_{HD}$]$_{OO}$]
SU PO VB
I think that you understand
'I think you understand.'
b. Jag vet
I know
[SuP varför$^{\boxed{1}}_{HD}$ [S hon$_{SB}$ kom$_{HD}$ hit.$_{RA}$ $\boxed{1}_{MD}$]$_{OO}$]
AB PO VB AB
why she came here
'I know what she came for.'

It is common for SuPs with a pseudo-head to be optionally or obligatorily doubly marked using the

[3] To overcome the limitations of the single line textual representation of structure, we use indexing for secondary edges: *node*$^{\boxed{i}}$ means that *node* will be referred to with index *i*, $\boxed{i}_{FN}$ means node *i* secondarily has function FN. The indices should not be understood as traces or null pronouns.

subordinator *som*, for instance when the pseudo-head is also subject in the complement S: *Ingen anar* [$_{\text{SuP}}$ *vad som* [$_{\text{S}}$ *sker.*]] 'No-one knows what goes on.'

Coordination KoP

Coordinations get their own phrase category, to deal with coordination of unlike categories.[4] The phrase category KoP can be understood as projected from the coordinator's part-of-speech KO. Coordinators are pseudoheads because of the existence of polysyndeton, in which head uniqueness is violated, (11).

(11) [$_{\text{KoP}}$ pappa och$_{\text{PH}}$ morfar och$_{\text{PH}}$ farfar]
NN KO NN KO NN
dad and grandpa and grandpa
'dad and grandpa (on mother's side) and grandpa (on father's side)'

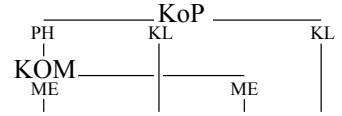
Next to subject sharing in the verbal domain, coordination is the other main application area for secondary annotations. They are used to distribute material over the conjuncts, as in (12).

(12) [$_{\text{NP}}$ en$^{\boxed{1}}_{\text{DT}}$ stuga] eller [$_{\text{NP}}$ $\boxed{1}_{\text{DT}}$ lada$_{\text{HD}}$]
PO NN NN
a cottage or barn
'a cottage or barn'

Multiword expressions *M

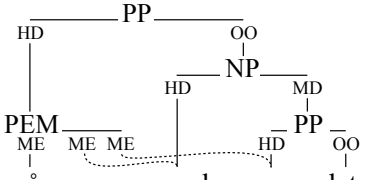
Multiword expressions are an important part of the Koala annotations, for two different reasons. First, in word sense annotation, multiword expressions as a whole will receive a single sense identifier from the SALDO lexicon. For singleword expressions, sense ids are attached to the token node, for multiword expressions, they are attached to a multiword node which connects to all elements of the expression using ME-labelled edges. Secondly, a part of the vocabulary of multiword expressions cannot be comfortably analyzed in syntactic terms using the general Koala schema – either because they show idiosyncratic properties or because they are part of expressions that can be said to have an expression specific grammar, for instance *Firstname Lastname* person names, street addresses, compound numerals, and so on.[5] We join all elements of such expressions directly under a (unstructured) multiword node, so that the whole may participate in the primary graph as if we were dealing with one token. On the one hand, this allows us to defer the question of whether such expressions should be one token or several (in terms of segmentation), on the other, it allows us to deal with a broader class of idiosyncratic expressions than a word-with-spaces approach, because material under a node need not be continuous. For instance, a discontinuous coordinator like *såväl ... som* 'both ... and' or a circumposition like *för ... sedan* 'ago' (lit. 'for ... since') is also gathered under one multiword node before participating in syntax as pseudo-head in a coordination or head in a PP.

Multiword expressions thus come in two flavours as far as Koala's annotation schema is concerned: analyzable and unanalyzable. Both types are annotated with the help of a multiword node to which we can attach a sense id. Unanalyzable multiword nodes have all their children in the primary graph. An example with a discontinuous coordinator is in (13).



(13) varken kött eller fisk
— NN KO NN
neither meat or fish
varken_eller..1
'neither meat nor fish'

Analyzable multiword expressions first receive a regular syntactic analysis, after which a multiword node is placed in the primary graph directly above one of the elements and the other elements are connected using secondary edges. The multiword annotation here solely fulfils the purpose of having a node to attach the SALDO annotation to. An example of a multiword preposition is given in (14).



(14) på grund av det
PE NN PE PO
on ground of that
på_grund_av..1
'because of that'

The analyzable multiwords can participate in syn-

[4] Note that, if needed, a more informative phrase type for the coordination can easily be derived automatically from the conjuncts in a coordination of like categories.

[5] This is not to say that the internal structure of such expressions is uninteresting.

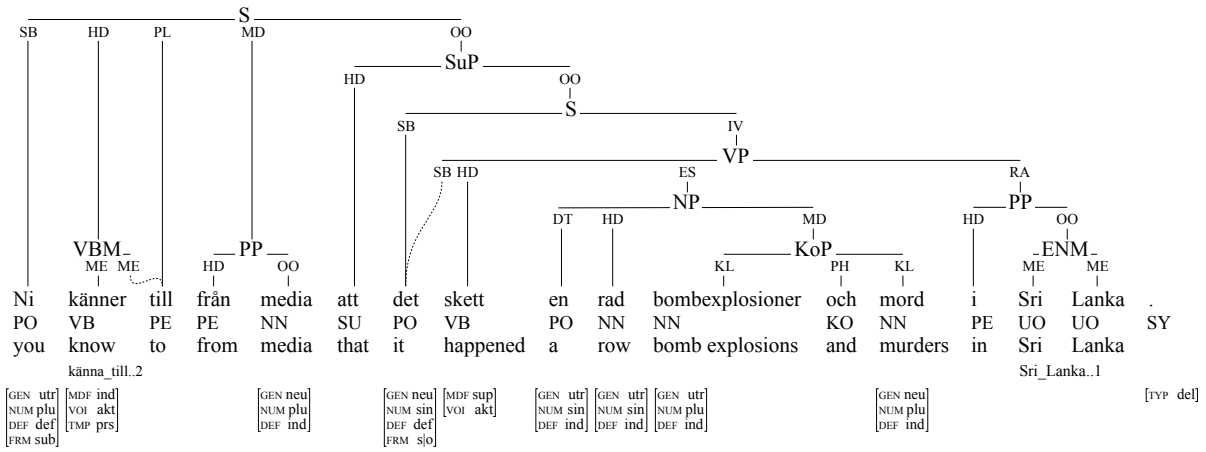


Figure 1: A full Koala sentence analysis.

tax to greater or lesser extent. Some, like the example in (14), are rather fixed, but others, like verb-object and -particle idioms, support-verb-constructions, etc., allow for more freedom, including modification of parts and flexible positioning of parts. Application of the distinction analyzable-unanalyzable has proven to be unproblematic for our annotators in practice, even though corner cases can be found.

4.4 A worked out example

We end this overview of Koala's morpho-syntactic annotation schema with a worked out complete example. Figure 1 shows the analysis of a sentence containing different types of subordinate clauses (S, SuP), two uses of secondary edges (in the multiword *känna till* 'know', lit. 'know to' vs subject control), two types of multiwords (the just mentioned vs *Sri Lanka*), so called *ha*-deletion (the missing temporal auxiliary governing the supine form *skett* 'happened'), a simple coordination, and a complex NP.

5 Conclusions

We have described the linguistic annotations of the 100k token mixed-genre Koala treebank, manually annotated with parts-of-speech and syntactic structures. The corpus will be freely available.

Both the inventory of parts-of-speech and the set of syntactic categories are more concise than in the de facto standards for annotating Swedish, SUC and MAMBA. This is because the simultaneous development of the two annotation levels has allowed us to carefully choose where to put which information. In particular, some part-of-speech distinctions that are purely based on function could be deferred to the syntactic level, with its hybrid structure of head-marked phrases and function labelled edges.

In addition, the structures should be easy to annotate, which means that the distinctions should be easy for the annotators to comprehend and apply. It also mean's that the structures are preferably compact: trees are relatively flat and do not contain empty nodes or unary nodes.

In contrast, we also want the syntactic structure to be easy to convert into other formalisms, which suggests a rich annotation. While the annotation is designed with an eye towards conversion into a bare constituency or dependency structure, we believe that the explicit annotation structure sharing and non-local relationships provided in the corpus can also make it usable as the basis for a conversion into linguistically richer formalisms (Cahill et al., 2004; Miyao et al., 2004).

Although the development of the annotation guidelines and the annotation itself is well underway, we have yet to do a thorough evaluation of the consistency of the annotation, the comprehensiveness of the annotation guidelines and the ease of annotating the described syntactic structures. However, at the time of writing we have annotations of parts-of-speech and syntactic structures for around 60k tokens. Our impression is that annotation is fast and the annotators enjoy the annotation work.

Acknowledgements

The Koala project is funded 2014–2016 by Riksbankens Jubilemsfond, grant number In13-0320:1.

References

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

Thorsten Brants, Wojciech Skut, and Hans Uszkoreit. 1999. Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop*, pages 69–76.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Aoife Cahill, Michael Burke, Ruth O'Donovan Josef Van Genabith, and Andy Way. 2004. Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 319–326.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*.

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

Helen Hoekstra, Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. 2001. Syntactic annotation for the spoken Dutch corpus project (CGN). In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000. Selected Papers from the Eleventh CLIN Meeting*, pages 73–87. Rodopi.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation.

Bengt Loman and Nils Jörgensen. 1971. *Manual för analys och beskrivning av makrosyntagmer*. Studentlitteratur, Lund.

Yusuke Miyao, Takashi Ninomiya, , and Jun'ichi Tsujii. 2004. Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the Penn Treebank. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP 2004)*, pages 684–693.

Joakim Nivre, 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.

Joakim Nivre, 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.

Joakim Nivre. 2002. What kinds of trees grow in Swedish soil? a comparison of four annotation schemes for Swedish. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for swedish. *Northern European Journal of Language Technology*, 3:1–18.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 88–95.

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

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur, Lund.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Tübingen.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim San Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. SMULTRON (version 3.0) — the Stockholm MULtilingual parallel TReebank. `http://www.cl.uzh.ch/research/parallelcorpora/paralleltreebanks_en.html`.