When Beards Start Shaving Men: A Subject-object Resolution Test Suite for Morpho-syntactic and Semantic Model Introspection

Patricia Fischer and **Daniël de Kok** and **Erhard Hinrichs**
University of Tübingen
{patricia.fischer, daniel.de-kok, erhard.hinrichs}
@uni-tuebingen.de

Abstract

In this paper, we introduce the SORTS **S**ubject-**O**bject **R**esolution **T**est **S**uite of German minimal sentence pairs for model introspection.[1] The full test suite consists of 18,502 transitive clauses with manual annotations of 8 word order patterns, 5 morphological and syntactic and 11 semantic property classes. The test suite has been constructed such that sentences are minimal pairs with respect to a property class. Each property has been selected with a particular focus on its effect on subject-object resolution, the second-most error-prone task within syntactic parsing of German after prepositional phrase attachment (Fischer et al., 2019). The size and detail of annotations make the test suite a valuable resource for natural language processing applications with syntactic and semantic tasks. We use dependency parsing to demonstrate how the test suite allows insights into the process of subject-object resolution. Based on the test suite annotations, word order and case syncretism can be identified as most important factors that affect subject-object resolution.

1 Introduction

Subject-object resolution remains a difficult task for syntactic parsing of languages with relatively free word order and case syncretism. For such languages, subject and object cannot always be disambiguated based on morpho-syntactic surface structures (Lenerz, 1977b; Eisenberg, 2013). Parser performance on German and Dutch, for example, has been shown to suffer significantly from incorrect identification of subject and object (Van Noord, 2007; Fischer et al., 2019). While there have been task-specific test suites for difficult syntactic phenomena in German such as prepositional phrase attachment, coordination, and verb phrase complementation (Nerbonne et al., 1991; Lehmann et al., 1996; Kübler et al., 2009), subject-object resolution has widely been neglected in existing test suites.

In this paper, we introduce the SORTS **S**ubject-**O**bject **R**esolution **T**est **S**uite of German minimal sentence pairs. The test suite has been created manually from hand-selected subject-verb-object triples and contains 18,502 transitive clauses with Universal Dependency (UD, De Marneffe et al. (2014)) annotations, template-based annotations of 8 word order patterns and manual annotations of 16 morphological, syntactic and semantic property classes. For instance, the sentence *Sie trifft eine Entscheidung.* 'She makes a decision.' has been annotated with word order `VF[S]LK[V]MF[O]`, the syntactic property `subject pronoun` and the semantic property `light verb` in addition to dependency relations relevant for subject-object resolution, as illustrated in Figure 1.

One domain of application for this test suite is syntactic parsing. We will use six neural dependency parsers with different architectures to show that the test suite is able to expose the linguistic properties that make subject-object resolution easier or more difficult for parsing. Experimental results imply that parsers are not able to resolve certain syntactic structures well, e.g. subject-object pairs with object-subject order and case syncretism.

The paper is structured as follows: Section 2 gives an overview of subject-object resolution in German as background to the test suite description in Section 3. Section 4 shows an application of the test suite

[1] The test suite is available online at https://github.com/DiveFish/SORTS and will continuously be extended. It is provided in CoNLL and sentence-based format, cf. Appendix A for samples from the test suite.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3019–3035
Barcelona, Spain (Online), December 8-13, 2020

in German dependency parsing. Related work is presented in Section 5 before concluding with some guiding remarks in Section 6.

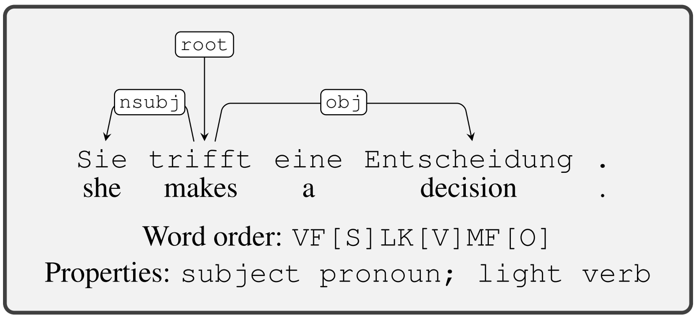
root
nsubj
obj
Sie trifft eine Entscheidung .
she makes a decision .
Word order: VF[S]LK[V]MF[O]
Properties: subject pronoun; light verb

Figure 1: Sample sentence from the SORTS test suite: annotation of subject and object dependency relations, word order, and additional linguistic properties.

2 Background

2.1 Subject-object Orders in German

In German, the order and position of subject and object are relatively flexible within a clause. As shown in Example 1 (adapted from TüBa-D/Z UD, sentence #35072), subject and object can be swapped without changing the meaning of the sentence (Lenerz, 1977b). Subjects and objects can be identified based on morphological case marking for masculine singular nouns, determiners and the majority of pronouns. In the remaining noun and determiner paradigms, nominative and accusative forms overlap. This case syncretism can result in ambiguities between subjects and direct objects as in Example 2. Here, both *das Ergebnis* and *Sie* can be nominative or accusative – in contrast to Example 1 where *euch* can only be accusative. Subject-verb agreement can resolve subject and object if subject and object differ in morphological marking for number.

(1) a. Erstaunt [das Ergebnis]$_{subj}$ [Euch]$_{obj}$? / Erstaunt [Euch]$_{obj}$ [das Ergebnis]$_{subj}$?
surprises [the result] [you-pl] ? / surprises [you-pl] [the result] ?

b. [Das Ergebnis]$_{subj}$ erstaunt [Euch]$_{obj}$. / [Euch]$_{obj}$ erstaunt [das Ergebnis]$_{subj}$.
[the result] surprises [you-pl] . / [you-pl] surprises [the result] .

(2) a. Erstaunt [das Ergebnis]$_{subj}$ [Sie]$_{obj}$? / Erstaunt [Sie]$_{obj}$ [das Ergebnis]$_{subj}$?
surprises [the result] [you-sg] ? / surprises [you-sg] [the result] ?

b. [Sie]$_{obj}$ erstaunt [das Ergebnis]$_{subj}$. / [Das Ergebnis]$_{subj}$ erstaunt [Sie]$_{obj}$.
[you-sg] surprises [the result] . / [the result] surprises [you-sg] .

Clause type	Topological fields	Example
Verb-first	$_{LK}$[V]$_{MF}$[SO] $_{LK}$[V]$_{MF}$[OS]	$_{LK}$[Erstaunt] $_{MF}$[das Ergebnis Sie]? $_{LK}$[Erstaunt] $_{MF}$[Sie das Ergebnis]? 'The result surprises you?'
Verb-second	$_{VF}$[S]$_{LK}$[V]$_{MF}$[O] $_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	$_{VF}$[Das Ergebnis] $_{LK}$[erstaunt] $_{MF}$[Sie]. $_{VF}$[Sie] $_{LK}$[erstaunt] $_{MF}$[das Ergebnis]. 'The result surprises you.'
Verb-last	$_{MF}$[SO]$_{VC}$[V] $_{MF}$[OS]$_{VC}$[V]	[Weil] $_{MF}$[das Ergebnis Sie] $_{VC}$[erstaunt]. [Weil] $_{MF}$[Sie das Ergebnis] $_{VC}$[erstaunt]. 'Because the result surprises you.'

Table 1: Topological field specifications for sentence *Erstaunt Sie das Ergebnis?* (TüBa-D/Z UD, sentence #35072).

German word order preferences such as the positions of subject and object have been described with the topological field model (Drach, 1937) that structures clauses in terms of fields: The middlefield (MF) is enclosed by the verbal bracket to the left (LK) and to the right (VC); the LK is preceded by the forefield (VF) in assertion main clauses.[2] The verb phrase is located in the LK and VC, subject and object in the VF or MF. The order of subject and object within the MF is relatively free. Table 1 lists examples of all subject-object orders and clause types specified with respect to topological fields.

2.2 Subject-object Resolution in Syntactic Parsing

Although subjects and objects can occur in any of the described positions within a German clause, a preference for subjects to precede objects can be observed. The TüBa-D/Z treebank of German newspaper text from the Berliner Tageszeitung *taz* with UD annotations (Telljohann et al., 2017; Çöltekin et al., 2017) includes 69,462 clauses with subject, verb and at least one object.[3] As Table 2 shows, the subject precedes the object in 81.31 percent of all clauses in the TüBa-D/Z UD.

The distribution of subject-object orders in corpora is also reflected by attachment scores of syntactic parsers trained on such corpora. Table 2 shows relative frequencies and labeled attachment score (LAS) per order of subject, verb and object in the TüBa-D/Z UD test set (12,657 samples). Attachment scores are based on the De Kok and Hinrichs (2016) parser, from here on "baseline parser". The baseline parser has been trained on 70 percent of the TüBa-D/Z UD with overall LAS 89.89. For the TüBa-D/Z UD test set, subject-object order frequency and LAS of the baseline parser positively correlate (Pearson correlation coefficient $\rho = 0.58$). More frequent word orders are parsed correctly more often. Consequently, subjects and objects are expected to be identified with lower accuracy for the less frequent object-subject order.

Word order	**Frequency in TüBa-D/Z (in %)**		**LAS**
	Full corpus	***Test set***	
$_{LK}$[V]$_{MF}$[SO]	13.70	13.15	87.60
$_{VF}$[S]$_{LK}$[V]$_{MF}$[O]	22.43	24.62	86.38
$_{MF}$[SO]$_{VC}$[V]	**45.09**	**45.05**	**89.52**
$_{LK}$[V]$_{MF}$[OS]	3.85	3.38	85.53
$_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	4.27	3.80	78.73
$_{MF}$[OS]$_{VC}$[V]	10.57	9.97	73.99

Table 2: Relative word order frequencies and LAS in the TüBa-D/Z UD (104,787 sentences and 1,959,474 tokens) and in the TüBa-D/Z UD test set (20,956 sentences and 390,129 tokens). In more than 80 percent of all clauses, the subject precedes the object. LAS positively correlates with relative order frequency (Pearson's $\rho = 0.58$).

3 The Test Suite

3.1 Setup

In German, there are three major types of clauses, which can be distinguished by the placement of the finite verb within the clause: verb-first, verb-second and verb-last clauses (Table 1). Besides verb-first and verb-last clauses, two kinds of verb-second clauses have been used in the SORTS test suite, one with one of the verbal arguments in the VF and one with an adverbial in the VF and both verbal arguments in the MF. Taking subject-object and object-subject orders into account, this results in 8 different word order patterns.[4]

[2] Abbreviations taken from the German terms *Vorfeld* (VF), *Linke Klammer* (LK), *Mittelfeld* (MF) and *Verbalkomplex* (VC).

[3] For clauses with more than one object, each object has been considered separately together with the subject and the verb.

[4] The number of test suite sentences with object-subject order is smaller than the number of subject-object sentences, with the exception of $_{VF}$[O]$_{LK}$[V]$_{MF}$[S]. The reason is that object-subject order in $_{VF}$[O]$_{LK}$[V]$_{MF}$[S] clauses is possible for any combination of subject and object whereas object-subject order within the MF allows only combinations where the object is more agentive than the subject (Lenerz, 1977a). For the exact number of sentences per property, cf. Table 7.

Clause type	Topological fields	Example
Verb-first	$_{LK}$[V]$_{MF}$[SO]	$_{LK}$[Erstaunt] $_{MF}$[das Ergebnis Sie]?
	$_{LK}$[V]$_{MF}$[OS]	$_{LK}$[Erstaunt] $_{MF}$[Sie das Ergebnis]?
		'The result surprises you?'
Verb-second	$_{VF}$[S]$_{LK}$[V]$_{MF}$[O]	$_{VF}$[Das Ergebnis] $_{LK}$[erstaunt] $_{MF}$[Sie].
	$_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	$_{VF}$[Sie] $_{LK}$[erstaunt] $_{MF}$[das Ergebnis].
		'The result surprises you.'
	$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[SO]	$_{VF}$[Deshalb] $_{LK}$[erstaunt] $_{MF}$[Sie das Ergebnis].
	$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[OS]	$_{VF}$[Deshalb] $_{LK}$[erstaunt] $_{MF}$[das Ergebnis Sie].
		'Therefore, the result surprises you.'
Verb-last	$_{MF}$[SO]$_{VC}$[V]	[Weil] $_{MF}$[das Ergebnis Sie] $_{VC}$[erstaunt].
	$_{MF}$[OS]$_{VC}$[V]	[Weil] $_{MF}$[Sie das Ergebnis] $_{VC}$[erstaunt].
		'Because the result surprises you.'

Table 3: Word order patterns in the SORTS test suite at the example of the sentence *Erstaunt Sie das Ergebnis?* (TüBa-D/Z UD, sentence #35072).

For each of the 8 patterns included in Table 3, we settle on a set of base sentences that we describe next. To each base sentence, we apply up to two of the morphological, syntactic, or semantic variations described in Sections 3.2 and 3.3. We take as the base case the subjects, objects and verbs which are found most frequently in the TüBa-D/Z UD and parsed most accurately by the baseline parser. The properties of these subjects, objects and verbs are the following:

- Subject and object noun phrases consist of a common noun preceded by a determiner. Table 4 shows that definite subjects occur most often with indefinite objects and are the easiest to parse of all definiteness combinations of subjects and objects. In order to avoid introducing biases of the linguistic expert creating the test suite sentences (first author of the paper), subjects, verbs and objects have been partially selected from frequent phrases in the TüBa-D/Z UD. Subjects and objects do not display case syncretism.

Subject	Object	Frequency (in %)	Subject-object LAS
Definite	Definite	40.63	83.72
	Indefinite	**59.37**	**86.27**
Indefinite	Definite	41.61	82.79
	Indefinite	58.39	85.62

Table 4: Relative frequency and subject-object LAS of definite and indefinite subjects and objects in the TüBa-D/Z UD test set. Definite subjects before indefinite objects are the most frequent subject-object combination with the highest LAS.

- Subjects have been restricted to be animate whereas objects are inanimate. Dowty (1991) describes the agent as prototypical subject of a clause. According to Eisenberg (2013), the degree of agency also decreases from subject to object. Since animacy is considered to correlate with agency, the subject has been defined to be animate and the object to be inanimate. One exception are psych verbs with experiencer objects as in *Das Auto gefällt ihnen.* 'The car pleases them.' where agency increases from subject to object (Lenerz, 1977b; Bader and Häussler, 2009).

- Unstressed object pronouns preferably occur at the left edge of the MF (Wackernagel position, Eisenberg (2013)). This affects the markedness of different subject-object orders: Sentences with pronominal object before a nominal (i.e. non-pronominal) subject will be less marked than a nominal object before a nominal subject. In order to avoid that different degrees of markedness influence subject-object resolution, subject and object have been defined not to be pronouns

- Verbs are in the active voice, do not include separable particles, modals or auxiliaries. The present tense has been used in all sentences. Furthermore, the chosen verbs take *accusative* objects as their argument. Monotransitive verbs with accusative object are by far more frequent than monotransitive verbs with dative objects (88.61 percent accusative compared to 8.50 percent dative objects in TüBa-D/Z UD). Accusative objects therefore allow a greater variety of verbs. Ditransitive verbs have been excluded in order to avoid that additional arguments introduce new sets of word order preferences which would in turn make a focused inspection of factors that affect subject-object resolution more difficult.

Subsections 3.2 and 3.3 introduce morphological, syntactic, and semantic properties that serve as the source of variation from the base case that is described in the present subsection.

3.2 Morphological and Syntactic Variations

Object case. The SORTS test suite contains monotransitive clauses with accusative and with dative objects. A comparison of clauses with accusative and dative objects can shed light on the effect of case on subject-object resolution.

Case syncretism. Case marking in German is not always unique. If both subject and object cannot be clearly identified based on their case and number markings, other cues to resolve subject and object need to be used. By enforcing case syncretism in parts of the test suite, it becomes possible to investigate how NLP systems deal with case syncretism. In combination with other selected properties, it can also be tested which properties make subject-object resolution for ambiguous subject-object sequences easier.

Pronominalization of subject and object. Pronouns have an effect on the preferred order of subject and object. In verb-last clauses with pronominal object, the object preferably occurs before a non-pronominal subject. Subject-object resolution should thus be easier for these clauses compared to other clauses where the object precedes the subject. If both subject and object are pronouns, the object cannot occur before the subject in verb-last clauses. Such sentences are considered ungrammatical and are therefore not part of the test suite.

Object negation. Results from experiments with the baseline parser showed that negated objects such as *keine Zeitung* 'no newspaper' are parsed correctly more often (LAS 95.19) than non-negated objects such as *eine Zeitung* 'a newspaper' (LAS 83.10). A variant with negated objects allows more focused investigations of the effect of negation on subject-object resolution.

Verb-argument distance: main verb position. Selectional restrictions originate from the main verb which selects its arguments such as the subject and the object. Making the verb phrase more complex by the help of an auxiliary verb changes the position of the main verb, increasing the distance between the main verb and its arguments. The auxiliary verb *werden* 'will' was used because it is not restricted to a particular class of verbs, in contrast to the auxiliary *haben* 'to have' or modal verbs such as *können* 'can/be able to'. Thus, *werden* serves as a means to test whether subject-object resolution becomes more difficult if the main verb and its arguments are further apart.

Verb-argument distance: additional constituents. As a seconds means to increase the distance between the main verb and its arguments, the temporal prepositional phrase (PP) *in dem Jahr* 'in that year' as the most frequent PP in the TüBa-D/Z UD[5] has been added in one sentence variant. The PP is inserted after the verb, if possible, as shown in Example 3. This avoids introducing further ambiguities such as PP attachment ambiguities.

(3)	Die	Leserin	abonniert	in	dem	Jahr	eine	Zeitschrift	.
	the	female-reader	subscribes-to	in	that	year	a	newspaper	.

[5] The PP pattern *in . . . Jahr* 'in . . . year' covers 0.77 percent of all PPs in the TüBa-D/Z UD.

3.3 Semantic Variations

Animacy. With respect to the semantic property of animacy, four co-occurrence patterns of subject and object have been included in the SORTS test suite: animate subject, inanimate object (base case pattern: no variation); inanimate subject, inanimate object (with deviation from case pattern: inanimate subject); animate subject, animate object (with deviation from base case pattern: animate object); inanimate subject, animate object (with deviation from the base case pattern for both subject and object). Suitable phrases were manually created, humans and other animal species being considered animate.

The decrease in agency from subject to object does not hold for **psych verbs with experiencer objects** such as *to amuse*, *to please* or *to frighten* where the object is typically more agentive than the subject (Lenerz, 1977b; Temme, 2018). For this reason, objects of psych verbs are animate, subjects inanimate, and psych verbs do not exhibit any of the variations for animacy described in the previous paragraph.

Regular polysemy. Languages tend to include nouns that exhibit regular patterns of polysemy (Apresjan, 1975). Nouns such as *university* and *company* can either refer to a set of people (in this reading: exhibiting the property of animacy) or a set of buildings (in this reading: lacking the animacy property). In order to determine whether a given occurrence of such nouns is more likely to be the subject or the object, we have to resolve the correct word sense of the noun. Instances of regular polysemy have been selected from the lexical-semantic word net GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). In all sentences, the animate readings have been used as subjects, since animate noun phrases are more agentive than inanimate noun phrases (cf. Section 3.1).

Proper name subjects. While common nouns are marked for case and number, proper nouns are marked for genitive case only. In addition, they are not accompanied by a determiner that could carry morphological information. In sum, proper nouns do not provide information about nominative or accusative and were therefore selected as more difficult cases of subject-object resolution.

Semantic asymmetry. Verbs can express actions that can only be executed between specific subjects and specific objects. The roles of subject and object cannot be swapped for these subject-verb-object combinations. One example is *to teach* with subject-object pairs such as *professor – student*. If not impossible, it is yet rare that *the student teaches the professor*. Other examples are *to command*, *to arrest* etc. The correct subject-object resolution is particularly challenging for such verbs since the correct asymmetric relation between the participants has to be established. Due to the lack of existing resources, examples of subject-object asymmetry have been manually created.

Non-referential objects. Non-referential objects such as inanimate *nichts* 'nothing' and animate *niemanden* 'nobody' make it possible to study the effects of object animacy and inanimacy on subject-object resolution, excluding other factors that may be due to the descriptive content of the head noun of a noun phrase.

Light verb constructions. Light verb constructions differ from other verb-object pairs in that the meaning is mostly derived from the object and the "light" verb contributes only little to the meaning of the phrase (Eisenberg, 2013). Examples are *eine Entscheidung treffen* 'to make a decision' or *eine Frage stellen* 'to pose a question'. Since the verb and the object form a unit of meaning, changes in word order should have less of an effect on subject-object resolution than with more loosely related verb-object pairs. Light verbs have been picked from the most frequent verb-object combinations in the TüBa-D/Z UD.

Verb synonyms. The meaning of a sentence as a whole is not affected by a replacement of the main verb by one of its synonyms. The same holds for the syntactic analysis of the sentence and its synonymous counterpart. If the syntactic analyses differ, the verb synonyms have erroneously been treated as two distinct, non-related verbs. Synonyms have been retrieved manually from Duden (2019) and an online thesaurus[6] by a German native speaker, aiming for minimal semantic differences between the original verb and its synonym.

[6] `https://synonyms.reverso.net/synonym-woerterbuch/` (last accessed 02 Oct 2020)

Idioms. In idioms, subject, object and verb form one unit of meaning (Duden, 2019). The variation of idioms into all different word orders allows insights into the importance of the syntactic structure of idioms. Idioms have been manually selected from a list of German idioms[7], including only idioms which consist of subject, verb and object without any additional phrases such as relative or coordinate clauses, adjuncts etc.

3.4 Degrees of Variation

Sentences with up to two variations from the base sentences, which were described in Section 3.1 above, were manually created for the SORTS test suite. Table 5 shows the different degrees of variation at an example sentence. These different degrees of variational depth make it possible to test if certain properties or property combinations make subject-object resolution easier or more difficult.

Variation	**Example**	**Property**
0: No variation	*Der Leser abonniert eine Zeitschrift.* 'the reader subscribes to a newspaper'	`base`
1: Auxiliary verb	*Der Leser <u>wird</u> eine Zeitschrift abonnieren.* 'the reader will subscribe to a newspaper'	`aux`
2: Auxiliary verb, synonymous verb	*Der Leser <u>wird</u> eine Zeitschrift <u>bestellen</u>.* 'the reader will order a newspaper'	`aux-syn`

Table 5: Examples for *Der Leser abonniert eine Zeitschrift.* in $_{VF}$[S]$_{LK}$[V]$_{MF}$[O] order.

3.5 Data Subsets

In order to determine the effect of ambiguity between subject and object on subject-object resolution, two data subsets have been created. In test suite SORTS$_{\text{part-amb}}$ (partially ambiguous), subject and object can be identified based on morphological information on subject and object, the only exception being one variation in which case syncretism occurs between subject and object (10,839 sentences). In test suite SORTS$_{\text{amb}}$ (fully ambiguous), all sentences have been manually changed to be ambiguous (7,663 sentences). In that set, the case syncretism variant has been removed along with the dative object variant for which no subject-object ambiguity occurs.

4 Parser Model Introspection

4.1 Parser Models

A neural transition-based dependency parser with a feed-forward neural network of one hidden layer serves as the baseline (De Kok and Hinrichs, 2016). In this Chen-and-Manning-(2014)-style parser, words and parts-of-speech are represented as structured skipgram embeddings (wang2vec, Ling et al. (2015)). In the case of word embeddings, subword units (Bojanowsky et al., 2017) are also used. Information about topological fields is provided as one-hot encodings.

The baseline parser has been extended in two ways: 1) Normalized PMIs from large corpora which indicate if and with which label a token should be attached to a candidate head, 2) similarity scores of dependency embeddings that have been trained to maximize the probability of two tokens occurring in a dependency relation in the training corpus (Fischer et al., 2019).

The simple feed-forward neural network of the baseline parser provides a very limited view of a token's context. The *sticker1* model builds on bidirectional LSTMs (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) which capture the left and right context of a token. The dependency edges are encoded as in Spoustová and Spousta (2010) and Strzyz et al. (2019). *Sticker1-self-distilled* makes additional use of self-distillation (Hinton et al., 2015; Furlanello et al., 2018). Both *sticker1* models use the same word embeddings as the baseline parser. *Sticker2* uses XLM-RoBERTa (Conneau et

[7] `https://de.wikiquote.org/wiki/Deutsche_Sprichw\%C3\%B6rter` (last accessed 02 Oct 2020)

al., 2019) finetuned on the TüBa-D/Z UD corpus for various morpho-syntactic tasks, including dependency parsing.[8] For consistent tokenization between all parsers, *sticker2* uses SentencePieces (Kudo and Richardson, 2018) on the token level.

4.2 Results and Error Analysis

The six parsers have been tested on the SORTS$_{\text{part-amb}}$ and SORTS$_{\text{amb}}$ test suites. Due to the focus on subject-object resolution, task-specific attachment scores are reported for subject and object heads and labels, discarding all other attachments.

Parser	**LAS SORTS$_{\text{no-amb}}$**	**LAS SORTS$_{\text{part-amb}}$**	**LAS SORTS$_{\text{amb}}$**
Subject-first	69.07	69.53	70.42
Baseline	82.44	81.90	70.44
Baseline + PMIs	82.56	82.06	70.69
Baseline + embeddings	82.32	81.88	70.79
sticker1	91.71	90.24	72.95
sticker1 + self-distilled	94.43	92.41	71.06
sticker2	**96.13**	**94.54**	**75.08**

Table 6: Subject-object LAS for test suites SORTS$_{\text{no-amb}}$, SORTS$_{\text{part-amb}}$ and SORTS$_{\text{amb}}$.

Table 6 shows subject-object LAS of all parsers for test suites SORTS$_{\text{part-amb}}$ and SORTS$_{\text{amb}}$. Results for SORTS$_{\text{part-amb}}$ are given including and excluding variants with case syncretism, the latter denoted as SORTS$_{\text{no-amb}}$ (no ambiguities). In addition to model LAS, scores for always choosing subject-object order in contrast to object-subject order are provided as *Subject-first*. Attachment scores decrease with a larger number of ambiguous sentences in the test suites just as improvements shrink over simply choosing the first verbal argument as the subject. *Sticker2* performs best on all test suites. It may benefit from having a deeper network (e.g. 12 layers compared to 3 layers in *sticker1-self-distilled*) and from pretraining on larger amounts of more varied data.

Property-specific results are given for the SORTS$_{\text{amb}}$ test suite. The lack of morphological indicators for subject-object identification makes the linguistic properties from Sections 3.2–3.3 more easily accessible than in SORTS$_{\text{part-amb}}$ (cf. Appendix C and D for subject-object LAS and baseline improvements for that test suite). Table 7 shows absolute baseline LAS and LAS improvements over the baseline parser for the five non-baseline parsers (cf. Appendix B for absolute LAS of all parsers). Linguistic properties are split into word order, morphological/syntactic and semantic properties. Results per property class represent task-specific subject-object LAS across all sentences to which the property applies. As suggested by the overall results, *sticker2* outperforms all other parsers by a wide margin for most of the linguistic properties. For SORTS$_{\text{part-amb}}$, *sticker2* achieves the best results on all but one property class with wider margins than for SORTS$_{\text{amb}}$.

In the word order category, more frequent subject-object orders show smaller improvements over the baseline than less frequent object-subject orders. One reason is that absolute LAS is already high for subject-object orders whereas results for object-subject orders range below 40.00 LAS points. As has been shown in Section 2.1, less frequent word orders are the most difficult for subject-object identification, in particular in combination with case syncretism between subject and object. However, these are also the cases where most can be gained from parsers with contextualized word representations that also take advantage of large amounts of training data. Results on the word order property in SORTS$_{\text{part-amb}}$ confirm these findings. For $_{MF}$[SO] orders, the larger number of parameters in the *sticker* models compared to the single feed-forward layer model of the baseline may explain the improvements over the baseline of *sticker1* and *sticker2*.

Within the morphological and syntactic category, *sticker2* improves considerably over the baseline for longer sentences with auxiliary verbs and PPs. On the semantic properties, *sticker2* outperforms other models in particular for proper name subjects and highly non-compositional clauses such as idioms.

[8] Software packages for *sticker1* and *sticker2*: `https://github.com/stickeritis`

For the latter, more training data seems to support more semantic representations of words and phrases which facilitates subject-object identification across different word orders. Low scores of *sticker2* for animate objects may originate from the fact that *sticker2* uses different word representations than the other parsers. *sticker2* uses the multilingual XLM-RoBERTa sentence piece vocabulary (Conneau et al., 2019) which consists of 250,000 pieces for more than 100 languages. The other parsers use a monolingual German word embedding vocabulary consisting of 710,288 words. Only one out of 34 animate objects is directly included in the vocabulary of the *sticker2* model which may have negative effects on the performance of *sticker2*. Similar data sparsity issues may be the reason for mixed results of the baseline parser with PMIs. Absolute subject-object LAS for proper name subjects and idioms confirm that *sticker2* improves the most on properties with relatively low baseline performance. Variations in performance between different properties for each individual parser underscore the utility of structuring the test suite according to word order, morphological, syntactic and semantic variations.

	Parser	**Baseline**	**Baseline**	**Baseline**	***sticker1***	***sticker1***	***sticker2***
Property (frequency)			**+PMIs**	**+embeds**		**+self-distill**	
Word order							
$_{LK}$[V]$_{MF}$[SO]	(1,349)	90.88	1.82	1.30	**3.71**	0.41	3.52
$_{VF}$[S]$_{LK}$[V]$_{MF}$[O]	(1,349)	95.00	0.63	**0.89**	-1.52	-5.78	-1.59
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[SO]	(1,349)	94.63	-0.07	-0.37	**0.89**	-3.78	0.48
$_{MF}$[SO]$_{VC}$[V]	(1,349)	88.14	-0.82	-0.56	**9.30**	8.12	7.64
$_{LK}$[V]$_{MF}$[OS]	(306)	1.47	-0.82	-0.33	0.98	0.33	**12.42**
$_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	(1,349)	29.91	-0.11	0.59	1.85	4.56	**10.04**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[OS]	(306)	0.82	-0.49	0.00	1.80	1.96	**10.46**
$_{MF}$[OS]$_{VC}$[V]	(306)	4.58	1.31	1.14	-2.45	-2.12	**4.90**
Morphological / syntactic							
No variation	(75)	91.33	2.00	3.33	-4.67	-2.67	**5.33**
Pronoun subject	(858)	82.58	1.05	0.99	1.69	0.17	**3.26**
Pronoun object	(1,213)	52.02	-0.29	-2.47	0.82	-1.03	**4.70**
Negated object	(1,175)	76.60	-2.04	0.43	2.68	-0.51	**3.57**
Auxiliary verb	(1,270)	72.76	0.67	-1.34	4.72	3.62	**5.75**
PP	(1,270)	65.47	1.93	1.34	8.11	12.72	**13.86**
Semantic							
Inanimate subject	(575)	79.91	2.00	3.48	4.96	3.22	**5.30**
Animate object	(715)	80.35	-1.68	**0.98**	-0.91	-6.15	-7.83
Inverted animacy	(811)	50.62	0.25	0.25	2.90	-3.95	**6.29**
Regular polysemy	(874)	77.23	-0.29	0.17	1.14	0.63	**3.83**
Proper name subject	(788)	64.34	0.70	1.33	4.51	7.49	**12.44**
Semantic asymmetry	(567)	74.43	0.97	**2.47**	1.41	-1.23	-0.79
Non-referential object	(783)	**77.01**	-2.23	-2.87	-4.53	-6.45	-9.07
Psych verb	(1,172)	46.54	0.17	0.85	1.92	-4.65	**7.85**
Light verb	(530)	94.15	0.19	-0.85	2.26	3.21	**3.87**
Synonymous verb	(1,150)	76.26	1.96	0.83	2.57	-0.74	**4.04**
Idiom	(155)	64.19	0.97	7.10	11.29	8.71	**12.90**

Table 7: Property-specific baseline subject-object LAS and baseline improvements on the SORTS$_{\text{amb}}$ test suite. The property frequency is included as the number of sentences per property.

5 Related Work

Grammar coverage and correctness. In rule-based parsing, parser performance was dependent on the correctness and coverage of grammar rules and the lexicon. Much work was devoted to testing and ensuring grammar coverage (Burkhardt, 1967; Purdom, 1972; Harrison et al., 1991). Error mining revealed weaknesses in the grammar by categorizing errors into classes that could be covered by adding new rules to the grammar (Van Noord, 2004; Sagot and De la Clergerie, 2006; De Kok et al., 2009).

Another approach to testing grammar coverage and correctness has been the design of test suites, with a noticeable interest in test suites on German syntax: Nerbonne et al. (1991) introduced a catalogue with the major syntactic patterns in German in order to facilitate error detection of NLP systems; Kübler et al. (2009) built a test suite for complex German constructions such as PP attachment, subject gaps and coordination of unlike constituents.

Test suite formats can range from unannotated lists of samples grouped by linguistic phenomena (Flickinger et al., 1987) to sets with detailed annotations such as *TSNLP – Test Suites for NLP* (Lehmann et al., 1996). Many previous attempts to create test suites were hampered by the diverse landscape of annotation schemes. The intent to make test suites more widely accessible led to efforts as the one by Kübler et al. (2009) who specifically provided test sentences in multiple annotation schemes. Universal Dependencies (De Marneffe et al., 2014) have successfully pushed the development of a language-independent annotation scheme. UD annotations have thus been used in the SORTS test suite. As they are not bound to a language-specific annotation scheme they make the test suite more applicable for international research in syntactic parsing and other fields where UD is now the de-facto standard annotation.

Probing. Increasing efforts are being made to investigate how linguistic structures are represented in neural networks. Linzen et al. (2016) probed an LSTM architecture's grammatical competence using training objectives with number prediction and grammaticality judgments in English as a target. Shwartz and Dagan (2019) present an evaluation suite consisting of tasks related to lexical composition, such as recognizing light verb and verb-particle constructions. Giulianelli et al. (2018) investigated how neural language models keep track of subject-verb agreement. Minimally-differing sentence pairs as they have been created for the SORTS test suite have also been used in the data sets by Poliak et al. (2018) and Ettinger et al. (2018).

6 Conclusion and Future Work

We presented the SORTS test suite for model introspection into subject-object resolution via German minimal sentence pairs. With a total size of 18,502 transitive clauses and 24 syntactic-semantic property classes, the test suite is a valuable resource for inspecting syntactic NLP systems of German. Its application in syntactic parsing revealed weaknesses of all parsers when syntactic and morphological cues are insufficient to resolve subjects and objects. Particularly difficult are sentences with object-subject order and case syncretism of subject and object. How to best resolve such sentences will remain an interesting case for future research on parsing.

Another direction for future work focuses on the extension of the test suite to cover more languages and linguistic phenomena. German still provides several cues to subject-object resolution: Case marking and subject-verb agreement. In Dutch, only some remainders of case marking persist in the pronoun paradigm whereas nominal subjects and objects can only be disambiguated morpho-syntactically through subject-verb agreement. For this reason, a second test suite is currently being constructed for Dutch. For ease of comparison, it will to a large extent be a translation of the German test suite.

Subject-object resolution is only the second largest error class behind PP attachment. A second test suite subset will be dealing with different aspects of PP attachment. In contrast to subject-object resolution, PP attachment has the advantage that it applies to a wider range of languages.

Acknowledgements

We would like to thank Prof. Dr. Christian Fortmann for his linguistic insights and valuable suggestions during the construction of the test suite. We also want to thank our reviewers for their constructive comments.

Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center "The Construction of Meaning" (SFB 833), project A3.

References

Jury Apresjan. 1975. Regular Polysemy. *Linguistics*, 142:5–32.

Markus Bader and Jana Häussler. 2009. Word Order in German: A Corpus Study. In *Lingua*, 120(3):717–762.

Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. In *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

W. H. Burkhardt. 1967. Generating Test Programs from Syntax. In *Computing*, 2(1):53–73.

Danqi Chen and Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750.

Çağrı Çöltekin, Ben Campbell, Erhard W. Hinrichs, and Heiek Telljohann. Converting the TüBa-D/Z Treebank of German to Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW)*, pages 27–37.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Javier Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale.

Daniël de Kok, Jianqiang Ma and Gertjan van Noord. 2009. A Generalized Method for Iterative Error Mining in Parsing Results. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks (GEAF)*, pages 71–79.

Daniël de Kok and Erhard W. Hinrichs. 2016. Transition-Based Dependency Parsing with Topological Fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1–7.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, JoakimNivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A Cross-Linguistic Typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. In *Language*, 67(3):547–619.

Erich Drach. 1937. *Grundgedanken der deutschen Satzlehre.* Diesterweg, Frankfurt am Main.

Duden (eds.). 2019. *Duden – Deutsches Universalwörterbuch.* Bibliographisches Institut, Berlin.

Peter Eisenberg. 2013. *Der Satz. Grundriss der deutschen Grammatik.* J.B. Metzler, Stuttgart.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing Composition Insentence Vector Representations. In *Proceedings of the 27th International Conference on Computational Linguistic (COLING)*, pages 1790–1801.

Patricia Fischer, Sebastian Pütz and Daniël de Kok. 2019. Association Metrics in Neural Transition-Based Dependency Parsing. In *Proceedings of the 5th International Conference on Dependency Linguistics (DepLing, SyntaxFest 2019)*, pages 181–189.

Daniel Flickinger, John Nerbonne, Ivan Sag, and Thomas Wasow. 1987. Towards Evaluation of Natural Language Processing Systems. *Technical report, Hewlett-Packard Laboratories.*

Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born Again Neural Networks.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the Hood: Using Diagnostic Classifiers to Investigate and Improve How Language Models Track Agreement Information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. In *Neural Networks*, 18(5-6):602–610.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Philip Harrison, Steven Abney, Ezra Black, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Don Hindle, Bob Ingria, Mitch Marcus, Beatrice Santorini, and Tomek Strzalkowski. 1991. Evaluating Syntax Performance of Parsers/Grammars of English. In *Proceedings of the ACL Workshop on Evaluating Natural Language Processing Systems*, pages 71–77.

Verena Henrich and Erhard W. Hinrichs. 2010. GernEdiT – the GermaNet Editing Tool. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. In *Neural Computation*, 9(8):1735–1780.

Sandra Kübler, Ines Rehbein, and Josef van Genabith (2009). TePaCoC – A Testsuite for Testing Parser Performance on Complex German Grammatical Constructions. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT).*

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language-Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP, System Demonstrations)*, pages 66–71.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Referik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. TSNLP – Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 711–716.

Jürgen Lenerz. 1977a. *Zum Einfluß von "Agens" auf die Wortstellung des Deutschen.* In *Grammatik und interdisziplinäre Bereiche der Linguistik. Akten des 11. Linguistischen Kolloquiums Aachen 1976* : 133–142.

Jürgen Lenerz. 1977b. *Zur Abfolge nominaler Satzglieder im Deutschen.* Gunter Narr, Tübingen.

Tal Linzen, Emmanuel Dupoux and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. In *Transactions of the Association for Computational Linguistics (TACL)*, 4:521–535.

Wang Ling, Chris Dyer, Alan Black and Isabel Trancoso. 2015. Two/Too Simple Adaptations of word2vec for Syntax Problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).*

John Nerbonne, Klaus Netter, Abdel Kader Diagne, Ludwig Dickmann, and Judith Klein. A Diagnostic Tool for German Syntax. In *Proceedings of the ACL Workshop on Evaluating Natural Language Processing Systems*, pages 79–96.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2227–2237.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White,and Benjamin Van Durme. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 67–81.

Paul Purdom. 1972. A Sentence Generator for Testing Parsers. In *BIT Numerical Mathematics*, 12(3):366-375.

Benoît Sagot and Éric de la Clergerie. 2006. Error Mining in Parsing Results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 329–336.

Vered Shwartz and Ido Dagan. 2019. Still a Pain in the Neck: Evaluating Text Representations on Lexical Composition. In *Transactions of the Association for Computational Linguistics (TACL)*, 7:403–419.

Drahomíra Spoustová and Miroslav Spousta. 2010. Dependency Parsing as a Sequence Labeling Task. In *The Prague Bulletin of Mathematical Linguistics*, 94(1):7–14.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Viable Dependency Parsing as Sequence Labeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2017. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).

Anne Temme. 2018. *The Peculiar Nature of Psych Verbs and Experiencer Object Structures.* Dissertation.

Gertjan van Noord. 2004. Error Mining for Wide-Coverage Grammar Engineering. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 446–463.

Gertjan van Noord. 2007. Using Self-Trained Bilexical Preferences to Improve Disambiguation Accuracy. In *Proceedings of the International Conference on Parsing Technologies (IWPT)*, pages 1–10.

Appendix A. Samples from the SORTS Test Suite

The SORTS test suite is delivered in two formats: the CoNLL format and the sentence-based format, which are shown below.

1) CoNLL format

The test suite in CoNLL format provides the sentences with each token being annotated for the different properties that apply for that sentence. Word order, marked by `order`, is separated from morphological, syntactic and semantic properties, marked by `props`. Head relations and head indices are given only for the tokens relevant for evaluation of subject-object resolution, namely subject, main verb and object. The example shows one of the sentences in the property combination of auxiliary verb and light verb construction `aux-vlight` in all possible word orders.

```
1  Die           _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        _  _      _ _
2  Chefin        _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        6  nsubj  _ _
3  wird          _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        _  _      _ _
4  eine          _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        _  _      _ _
5  Entscheidung  _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        6  obj    _ _
6  treffen       _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        0  verb   _ _
7  .             _ _ _  order:VF[S]LK[V]MF[O]|props:aux-vlight        _  _      _ _

1  Eine          _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        _  _      _ _
2  Entscheidung  _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        6  obj    _ _
3  wird          _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        _  _      _ _
4  die           _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        _  _      _ _
5  Chefin        _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        6  nsubj  _ _
6  treffen       _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        0  verb   _ _
7  .             _ _ _  order:VF[O]LK[V]MF[S]|props:aux-vlight        _  _      _ _

1  Deshalb       _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     _  _      _ _
2  wird          _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     _  _      _ _
3  die           _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     _  _      _ _
4  Chefin        _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     7  nsubj  _ _
5  eine          _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     _  _      _ _
6  Entscheidung  _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     7  obj    _ _
7  treffen       _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     0  verb   _ _
8  .             _ _ _  order:VF[ADV]LK[V]MF[SO]|props:aux-vlight     _  _      _ _

1  Wird          _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           _  _      _ _
2  die           _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           _  _      _ _
3  Chefin        _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           6  nsubj  _ _
4  eine          _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           _  _      _ _
5  Entscheidung  _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           6  obj    _ _
6  treffen       _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           0  verb   _ _
7  ?             _ _ _  order:LK[V]MF[SO]Q|props:aux-vlight           _  _      _ _
```

1	Weil	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	_	_	_	_
2	die	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	_	_	_	_
3	Chefin	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	6	nsubj	_	_
4	eine	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	_	_	_	_
5	Entscheidung	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	6	obj	_	_
6	treffen	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	0	verb	_	_
7	wird	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	_	_	_	_
8	.	_	_	_	order:MF[SO]VC[V]|props:aux-vlight	_	_	_	_

2) Sentence-based format

The test suite in sentence-based format provides the sentences with each sentence being annotated for the different properties that apply for that sentence. Word order in column 1 is separated from morphological, syntactic and semantic properties in column 2. Subject and object indices for easy subject-object identification are given in column 3 and 4. This format is particularly suited for architectures which take as input full sentences. Again, the example shows one of the sentences in the property combination of auxiliary verb and light verb construction `aux-vlight` in all possible word orders.

Word order	**Properties**	**S**	**O**	**Sentence**
VF[S]LK[V]MF[O]	aux-vlight	2	5	Die Chefin wird eine Entscheidung treffen .
VF[O]LK[V]MF[S]	aux-vlight	5	2	Eine Entscheidung wird die Chefin treffen .
VF[ADV]LK[V]MF[SO]	aux-vlight	4	6	Deshalb wird die Chefin eine Entscheidung treffen .
LK[V]MF[SO]Q	aux-vlight	3	5	Wird die Chefin eine Entscheidung treffen ?
MF[SO]VC[V]	aux-vlight	3	5	Weil die Chefin eine Entscheidung treffen wird .

Appendix B. Subject-object LAS on SORTS$_{\text{amb}}$

Parser Property	Baseline	Baseline +PMIs	Baseline +embeds	*sticker1*	*sticker1* +self-distill	*sticker2*
Word order						
$_{LK}$[V]$_{MF}$[SO]	90.88	92.70	92.18	**94.59**	91.29	94.40
$_{VF}$[S]$_{LK}$[V]$_{MF}$[O]	95.00	95.63	**95.89**	93.48	89.21	93.40
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[SO]	94.63	94.55	94.26	**95.52**	90.85	95.11
$_{MF}$[SO]$_{VC}$[V]	88.14	87.32	87.58	**97.44**	96.26	95.77
$_{LK}$[V]$_{MF}$[OS]	1.47	0.65	1.14	2.45	1.80	**13.89**
$_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	29.91	29.80	30.50	31.76	34.47	**39.96**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[OS]	0.82	0.33	0.82	2.61	2.78	**11.27**
$_{MF}$[OS]$_{VC}$[V]	4.58	5.88	5.72	2.12	2.45	**9.48**
Morpho-syntactic						
No variation	91.33	93.33	94.67	86.67	88.67	**96.67**
Pronoun subject	82.58	83.62	83.57	84.27	82.75	**85.84**
Pronoun object	52.02	51.73	49.55	52.84	50.99	**56.72**
Negated object	76.60	74.55	77.02	79.28	76.09	**80.17**
Auxiliary verb	72.76	73.43	71.42	77.48	76.38	**78.50**
PP	65.47	67.40	66.81	73.58	78.19	**79.33**
Semantic						
Inanimate subject	79.91	81.91	83.39	84.87	83.13	**85.22**
Animate object	80.35	78.67	**81.33**	79.44	74.20	72.52
Inverted animacy	50.62	50.86	50.86	53.51	46.67	**56.91**
Regular polysemy	77.23	76.95	77.40	78.38	77.86	**81.06**
Proper name subject	64.34	65.04	65.67	68.85	71.83	**76.78**
Semantic asymmetry	74.43	75.40	**76.90**	75.84	73.19	73.63
Non-referential object	**77.01**	74.78	74.14	72.48	70.56	67.94
Psych verb	46.54	46.72	47.40	48.46	41.89	**54.39**
Light verb	94.15	94.34	93.30	96.42	97.36	**98.02**
Synonymous verb	76.26	78.22	77.09	78.83	75.52	**80.30**
Idiom	64.19	65.16	71.29	75.48	72.90	**77.10**

Table 8: Property-specific subject-object LAS on the SORTS$_{\text{amb}}$ test suite.

Appendix C. Baseline improvements on SORTS$_{\text{part-amb}}$

Parser **Property (frequency)**	**Baseline +PMIs**	**Baseline +embeds**	***sticker1***	***sticker1* +self-distill**	***sticker2***
Word order					
$_{LK}$[V]$_{MF}$[SO] (1,884)	0.50	0.45	5.36	6.00	**7.19**
$_{VF}$[S]$_{LK}$[V]$_{MF}$[O] (1,884)	0.32	0.45	0.93	0.08	**1.33**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[SO] (1,884)	-0.96	-0.93	3.05	3.48	**4.78**
$_{MF}$[SO]$_{VC}$[V] (1,884)	0.50	0.00	5.47	5.89	**6.29**
$_{LK}$[V]$_{MF}$[OS] (473)	5.60	0.85	18.92	27.70	**35.73**
$_{VF}$[O]$_{LK}$[V]$_{MF}$[S] (1,884)	-0.13	1.22	14.68	19.59	**23.81**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[OS] (473)	0.95	-4.44	22.41	34.57	**38.58**
$_{MF}$[OS]$_{VC}$[V] (473)	-3.70	-1.59	32.45	39.01	**42.60**
Morpho-syntactic					
No variation (75)	-1.33	0.00	2.67	2.67	**2.67**
Dative object (1,285)	-0.62	-1.48	6.69	7.78	**9.38**
Case syncretism (1,335)	0.49	0.71	1.72	-0.04	**5.17**
Pronoun subject (1,140)	2.81	2.68	7.72	8.77	**9.56**
Pronoun object (1,387)	-1.23	-1.80	12.44	15.93	**17.09**
Negated object (1,405)	-0.07	1.07	7.65	8.75	**11.39**
Auxiliary verb (1,467)	0.03	-2.42	11.69	12.41	**13.33**
PP (1,490)	2.01	1.78	17.99	26.51	**27.89**
Semantic					
Inanimate subject (812)	-1.17	0.49	6.22	8.99	**10.10**
Animate object (870)	1.15	1.38	3.16	4.02	**5.11**
Inverted animacy (1,182)	0.13	-1.18	14.89	20.73	**22.67**
Regular polysemy (1,065)	-0.28	-0.33	4.32	5.92	**8.50**
Proper name subject (945)	0.58	0.42	5.61	8.94	**12.96**
Semantic asymmetry (720)	-0.83	0.28	3.54	**4.72**	3.96
Non-referential object (1,065)	-2.72	-2.82	7.32	9.81	**10.23**
Psych verb (1,560)	1.19	0.64	14.74	16.79	**22.98**
Light verb (675)	1.41	0.89	4.30	5.04	**5.41**
Synonymous verb (1,320)	0.04	-0.76	4.81	7.27	**8.18**
Idiom (320)	0.63	5.63	10.00	7.19	**15.78**

Table 9: Property-specific baseline improvements in subject-object LAS on the SORTS$_{\text{part-amb}}$ test suite. The property frequency is included as the number of sentences per property.

Appendix D. Subject-object LAS on SORTS$_{\text{part-amb}}$

Parser	**Baseline**	**Baseline**	**Baseline**	***sticker1***	***sticker1***	***sticker2***
Property		**+PMIs**	**+embeds**		**+self-distill**	
Word order						
$_{LK}$[V]$_{MF}$[SO]	91.06	91.56	91.51	96.42	97.05	**98.25**
$_{VF}$[S]$_{LK}$[V]$_{MF}$[O]	97.66	97.98	98.12	98.59	97.74	**98.99**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[SO]	93.55	92.60	92.62	96.60	97.03	**98.33**
$_{MF}$[SO]$_{VC}$[V]	92.28	92.78	92.28	97.74	98.17	**98.57**
$_{LK}$[V]$_{MF}$[OS]	44.40	50.00	45.24	63.32	72.09	**80.13**
$_{VF}$[O]$_{LK}$[V]$_{MF}$[S]	65.10	64.97	66.32	79.78	84.69	**88.91**
$_{VF}$[ADV]$_{LK}$[V]$_{MF}$[OS]	43.34	44.29	38.90	65.75	77.91	**81.92**
$_{MF}$[OS]$_{VC}$[V]	37.84	34.14	36.26	70.30	76.85	**80.44**
Morpho-syntactic						
No variation	97.33	96.00	97.33	100.00	100.00	**100.00**
Dative object	87.55	86.93	86.07	94.24	95.33	**96.93**
Case syncretism	78.05	78.54	78.76	79.78	78.01	**83.22**
Pronoun subject	88.11	90.92	90.79	95.83	96.89	**97.68**
Pronoun object	76.68	75.45	74.87	89.11	92.61	**93.76**
Negated object	84.52	84.45	85.59	92.17	93.27	**95.91**
Auxiliary verb	84.08	84.12	81.66	95.77	96.49	**97.41**
PP	67.42	69.43	69.19	85.40	93.93	**95.30**
Semantic						
Inanimate subject	87.56	86.39	88.05	93.78	96.55	**97.66**
Animate object	91.67	92.82	93.05	94.83	95.69	**96.78**
Inverted animacy	70.09	70.22	68.91	84.98	90.82	**92.77**
Regular polysemy	87.84	87.56	87.51	92.16	93.76	**96.34**
Proper name subject	75.40	75.98	75.82	81.01	84.34	**88.36**
Semantic asymmetry	90.97	90.14	91.25	94.51	**95.69**	94.93
Non-referential object	88.87	86.15	86.06	96.20	98.69	**99.11**
Psych verb	65.26	66.44	65.90	80.00	82.05	**88.24**
Light verb	92.74	94.15	93.63	97.04	97.78	**98.15**
Synonymous verb	88.48	88.52	87.73	93.30	95.76	**96.67**
Idiom	73.75	74.38	79.38	83.75	80.94	**89.53**

Table 10: Property-specific subject-object LAS on the SORTS$_{\text{part-amb}}$ test suite.