The Universal Dependencies Treebank of Spoken Slovenian

Kaja Dobrovoljc[1]**, Joakim Nivre**[2]
[1]Institute for Applied Slovene Studies Trojina, Ljubljana, Slovenia
[1]Department of Slovenian Studies, Faculty of Arts, University of Ljubljana
[2]Department of Linguistics and Philology, Uppsala University
kaja.dobrovoljc@trojina.si, joakim.nivre@lingfil.uu.se

Abstract

This paper presents the construction of an open-source dependency treebank of spoken Slovenian, the first syntactically annotated collection of spontaneous speech in Slovenian. The treebank has been manually annotated using the Universal Dependencies annotation scheme, a one-layer syntactic annotation scheme with a high degree of cross-modality, cross-framework and cross-language interoperability. In this original application of the scheme to spoken language transcripts, we address a wide spectrum of syntactic particularities in speech, either by extending the scope of application of existing universal labels or by proposing new speech-specific extensions. The initial analysis of the resulting treebank and its comparison with the written Slovenian UD treebank confirms significant syntactic differences between the two language modalities, with spoken data consisting of shorter and more elliptic sentences, less and simpler nominal phrases, and more relations marking disfluencies, interaction, deixis and modality.

Keywords: dependency treebank, spontaneous speech, Universal Dependencies

1. Introduction

It is nowadays a well-established fact that data-driven parsing systems used in different speech-processing applications benefit from learning on annotated spoken data, rather than using models built on written language observation. Since the influential syntactic annotation of the Switchboard section of the Penn Treebank (Godfrey et al., 1992; Marcus et al., 1993), several syntactically annotated spoken corpora have emerged, such as the Verbmobil treebanks for English, German and Japanese (Hinrichs et al., 2000), the CGN treebank for Dutch (van der Wouden et al., 2002), the NoTa treebank for Norwegian (Johannessen and Jørgensen, 2006), the PDTSL treebank for Czech (Hajič et al., 2008), and the Rhapsodie treebank for French (Lacheret et al., 2014). However, until now, no syntactically annotated data has been available for spoken Slovenian.

In addition to differences in the underlying phrase-based or dependency-based grammatical formalisms, existing spoken treebanks vary considerably in their approach to annotation of syntactic particularities of spoken language, even though these are not generally considered as language-specific. On one side of the spectrum are annotation schemes providing syntactic analysis of all transcribed lexical tokens, typically by introduction of new labels for speech-specific phenomena, while on the other side we find schemes, in which only well-formed, written-like constructions are included in the resulting syntactic trees, disregarding disfluencies and other types of 'noisy' structural particularities.

This prevalent multi-layer approach has partially been motivated by the data-driven parsing systems themselves, usually adopting a two-pass pipeline architecture, in which the structural particularities are first removed and followed by parsing of normalized transcriptions (Charniak and Johnson, 2001). Recent advances in parsing systems using non-monotonic transition-based algorithms, however, show that joint treatment of disfluencies and other syntactic relations actually out-performs state-of-the-art pipeline approaches (Rasooli and Tetreault, 2013; Honnibal and Johnson, 2014). Such heterogeneity of spoken language annotation schemes inevitably leads to a restricted usage of existing spoken language treebanks in linguistic research and parsing systems alike, limiting any direct comparison between spoken language treebanks of different formalisms, modalities (spoken or written) or languages. The need for a cross-linguistically harmonized treatment of non-language-specific phenomena is even more important in the field of spoken language resources, as these are still very limited in terms of number, size and availability due to their costly construction.

To ensure its wide and long-term usability, the new treebank of spoken Slovenian adopts the recently proposed Universal Dependencies annotation scheme, aimed at cross-linguistically consistent dependency treebank annotation. In the following part of this paper, we first briefly describe the process of the treebank construction and the general principles related to its tokenization, segmentation and spelling. Given this is the first attempt to apply the Universal Dependencies scheme to extensive spoken data, we then present its adaptation for various types of speech-specific phenomena, describe the annotation process, and show how the new spoken treebank compares to its written counterpart.

2. Treebank Construction

The Spoken Slovenian Treebank is a sample of the Gos reference corpus of Spoken Slovenian (Zwitter Vitez et al., 2013), a collection of audio recordings and transcripts of approximately 120 hours (1 million words) of monologic, dialogic and multi-party spontaneous speech in different everyday situations. The reference corpus was balanced to be representative of speakers (sex, age, region, education), communication channels (TV, radio, telephone, personal contact) and spoken communication settings, broadly categorized into *public informative and educational* (TV and radio shows, interviews, debates; school lessons, academic lectures), *public entertainment* (talk shows, morning radio shows, sports broadcasting), *non-public non-private* (work meetings, consultations, sale and other services) and *non-public private* (conversations between friends or family).

To ensure a similar distribution of text type, channel and speaker demographics to the reference corpus, the Spoken Slovenian Treebank was sampled by taking a random segment with a proportional number of tokens from each of the 287 texts in the original corpus. Each sampled text segment consists of one or more subsequent *turns* (units of speech by one speaker), which in themselves consist of one or more *utterances* (semantically, syntactically and acoustically delimited units, roughly corresponding to written sentences).[1].

Thus, the large majority of texts in the treebank include longer continuous spans of complete turns by one or more speakers, enabling posterior extension, re-segmentation or addition of other layers of linguistic annotation, such as discourse relations annotation or dialogue act annotation. A detailed description of the sampled treebank, currently amounting to 3,188 utterances or 29,468 tokens, is given in Table 1.

Type	Texts	Speak.	Turns	Utter.	Tokens
PI	129	263	703	959	9,898
PE	42	78	499	726	6,826
NN	45	102	425	497	4,535
NP	71	163	833	1,006	8,209
Total	287	606	2,460	3,188	29,468

Table 1: Treebank size by text type: PI = public informative and educational; PE = public entertainment; NN = non-public non-private; NP = non-public private.

3. Segmentation, Tokenization and Spelling

Typically, spoken language annotation denotes annotation of its representation in the form of written transcription. In the Spoken Slovenian Treebank, the spelling, tokenization and segmentation principles follow the transcription guidelines of the reference Gos corpus (Verdonik et al., 2013). The syntactic trees in the treebank span over individual utterances, manually delimited in the process of reference corpus transcription. Nevertheless, given the notoriously difficult task of speech segmentation, the turn-based sampling of the treebank enables posterior re-segmentation of utterances into longer or shorter units, if necessary.

Among the two types of Gos transcriptions (pronunciation-based and normalized spelling, both in lowercase only), the morphological and syntactic analysis is performed on top of normalized transcriptions that reduce the number of token types due to regional, colloquial and other pronunciation variation. In case of tokenization discrepancies between the two levels of transcription (e.g. between the colloquial *nauš* 'you won't' and the normalized *ne boš* 'you will not'), the normalized tokenization is selected, but the mapping of both spellings is maintained.

In addition to lexical tokens (words), the transcripts also include tokens signalling filled pauses (fillers), unfinished or incomprehensible words, as well as extralinguistic tokens marking basic prosodic information, such as exclamation or interrogation intonation markers, silent pauses (if longer than 1.5 sec), non-turn taking speaker interruptions, vocal sounds (e.g. laughing, sighing, yawning) and non-vocal sounds (e.g. applauding, ringing). In the treebank, *all* transcription tokens are considered nodes of dependency trees, however, it is a straightforward task to filter out the non-lexical tokens and obtain representations with words only.

4. Annotation Scheme

4.1. Universal Dependencies

Universal Dependencies[2] is a recently proposed annotation scheme for development of cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective (Nivre, 2015). It is the result of previous similar standardization projects (Zeman, 2008; Petrov et al., 2012; Marneffe et al., 2014) and has already been applied to more than 30 different languages (Nivre et al., 2015), including (written) Slovenian. A detailed description of the design principles and the relation taxonomy is given in Nivre (2015) and Nivre et al. (2016), with the main principles being that dependency relations hold primarily between content words, function words attach to the content word they specify and punctuation marks attach to the clause or phrase to which they belong. The *basic* dependency representation forms a tree, but additional dependencies can be added in the so-called *enhanced* representation.

From the perspective of spoken language annotation, the two most important features are that the universal taxonomy already includes labels for several speech-specific loose-joining syntactic relations, such as *reparandum*, *parataxis*, *discourse*, *dislocated*, and *vocative*, and that the scheme design allows for language-specific extensions, when necessary. Although metadata for UD release v1.2 shows that treebanks for Estonian, Greek, Danish and Persian also include some spoken texts, the lack of mention of this modality in the corresponding treebank documentations suggests this is the first systematic application of the Universal Dependencies scheme to extensive amount of spoken data.

4.2. Adaptation to Speech

As we have already emphasized in the introduction section, our ambition is to account for all syntactic phenomena in speech in a unified dependency annotation scheme, in which all tokens are treated as dependents belonging to the same syntactic tree. In the following section, we present Universal Dependency annotation principles for the most frequent speech-specific constructions, addressing both structural ('malformed') and pragmatic ('well-formed') particularities. The necessary adaptations of the

[1] The only exception are monologue-like academic lectures, in which turns by one speaker often span over entire speech events, so sampling of an incomplete turn was allowed (5.5% of all sampled turns)

[2] http://universaldependencies.org/

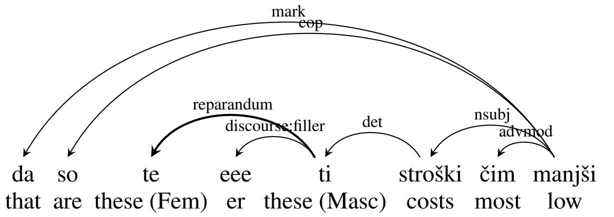
scheme are done either by extending the scope of application of an existing universal label or by introducing a new speech-specific extension, represented as a concatenation of the governing universal label, a colon mark and the extension sub-label, e.g. *label:sub-label*.

4.2.1. Extralinguistic Tokens

The Spoken Slovenian Treebank includes the following groups of extralinguistic tokens: markers of interrogative (`?`) and exclamatory (`!`) intonation, markers of silent pauses (`[pause]`), markers of unfinished or incomprehensible speech fragments (`[gap]`), markers of vocal (`[:voice]`) and non-vocal sounds (`[incident]`), markers of laughter (`[speaker:laughter]`, `[audience:laughter]`, `[all:laughter]`), and markers of non-turn taking speaker interruptions (`...`). These tokens, tagged as *X* or *PUNCT* (?, !, ...), are labeled as *punct* on the dependency layer and follow the general UD principle of attaching punctuation nodes to the highest relevant node that preserves projectivity.

4.2.2. Disfluencies

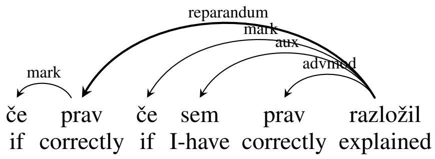
On the general level, our annotation scheme distinguishes two different types of syntactic disfluencies in speech: repairs and restarts. Repairs are instances where an edited unit (the reparandum) gets overridden by a new unit (the repair). Reparandums can be edited either by repetition of the same word or strings of words, usually to gain planning time, by substitution with a different word form, usually to correct its grammatical features (Figure 1), or by reformulation with a new word or phrase, to modify the intended verbalisation (Figure 2). Regardless of the type of repair (repetition, substitution, reformulation), and the type of reparandum (complete or incomplete; word fragment, word or phrase), the head of the edited unit is labeled as *reparandum* and depends on its repair.[3]


mark
cop
reparandum
discourse:filler
det
nsubj
advmod
da so te eee ti stroški čim manjši
that are these (Fem) er these (Masc) costs most low

(so that these costs are as low as possible)

Figure 1: Annotation of substituted units by *reparandum*.

In cases when the edited unit is a syntactically incomplete string of nodes (as in the example in Figure 2, we follow the principle of head promotion in ellipsis (see the following section), and attach the reparandum to the highest node of the following unit that preserves projectivity. This includes repairs of speech fragments (otherwise tagged as *X*).


reparandum
mark
aux
advmod
mark
če prav če sem prav razložil
if correctly if I-have correctly explained

(if I explained correctly)

Figure 2: Annotation of syntactically incomplete edited units by *reparandum*.

The second type of self-editing disfluencies are restarts (termed deletions by Shriberg (1996)), in which an unfinished clause is abandoned and replaced by a new sentence, as in the example in Figure 3. We label such paratactical restarts with a speech-specific label *parataxis:restart* that spans from the predicate head of the unfinished clause.

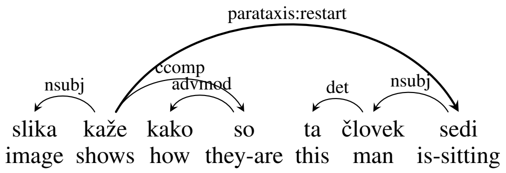

parataxis:restart
ccomp
advmod
nsubj
det
nsubj
slika kaže kako so ta človek sedi
image shows how they-are this man is-sitting

Figure 3: Annotation of abandoned sentences by *parataxis:restart*.

Given the relatively arbitrary position of interruption points, at which a phrase or a clause is being abandoned and replaced by its edit, the edited unit is therefore often left syntactically incomplete, as in Figures 2 and 3. The randomly elliptical nature of trees involving the *reparandum* and *parataxis:restart* labels should thus be given appropriate attention in both the process of parser training and evaluation. Similarly, the intended syntactic function of an incomplete reparandum unit, such as *advmod* in Figure 2 cannot be directly inherited from its repair (as with a syntactically complete reparandum in Figure 1), but explicit relations or attachments to nodes outside the repair could later be added as part of the enhanced representation.

4.2.3. Ellipsis

In addition to ellipsis due to self-editing, interruptions or unfinished utterances, there are also many instances of predicate ellipsis due to inference from the textual or situational context, such as in continuation of a topic (*pri nas pa občasno* ('we (do it) only once in a while'), formulaic answers or replies (*kam pa?* 'where to?', *zakaj ne?* 'why not?', *upam da* 'I hope so'), introductions to reported speeech (*in pol jaz njemu* 'and then I (say) to him'), sports commentaries (*tudi francozinja v težavah* 'the French (is) also in troubles'), etc. Following the general UD guidelines for treatment of ellipsis, the orphan node gets promoted to the place of the missing parent. In case of several same-level dependents of the missing head, elements of the verb phrase are given priority over other constituents, content

[3] For syntactically complete reparandum units, we currently consider its repair to be the equivalent syntactic unit, even if the repair itself is a constituent of larger units. We thus give preference to syntactic equivalence, e.g. between repaired modifiers or function words, rather than to high attachment.

words over function words, core arguments over non-core arguments, propositional adverbials (adjuncts) over non-propositional (disjuncts), and so on, as in the examples in Figures 2 and 3, in which the adverb and the auxiliary have been promoted to the position of the missing main verb.

4.2.4. Discourse Elements

On the lexical level, spoken communication includes many words and expressions that do not contribute much to the propositional content of what is being said, but instead function on some level of discourse organisation, as explicit markers of discourse relations, discourse structure, interaction management, speaker attitudes, etc. Different theories define and categorize these expressions differently, with often overlapping terminology (such as discourse connectives, discourse/modal particles, discourse/pragmatic markers, etc.) and the acknowledgment of fuzzy boundaries between them (Degand et al., 2013; Fischer, 2006). The highly multifunctional nature of discourse-related lexica therefore renders its part-of-speech and syntactic categorisation particularly difficult.[4]

In the Spoken Slovenian Treebank, three groups of discourse-related problematic expressions have been identified and addressed. For non-relational discourse elements, such as interjections (*aha* 'uh-huh', *opsala* 'oops'), response tokens (*ja* 'yes', *itak* 'sure'), expressions of politeness (*hvala* 'thanks', *adijo* 'bye') and prototypical non-clausal discourse markers (*no* 'well', *zdaj* 'now', *a ne* 'right?'), we use the *discourse* label, and a newly introduced extension *discourse:filler* for filler sounds (*eee* 'uh'). Both labels attach to the highest relevant unit preserving projectivity, which can either be a predicate, as in the case of discourse markers or response tokens (Figure 4), or any other constituent, as in the case of repairs, where fillers or other editing signals attach to the repair (Figure 1).

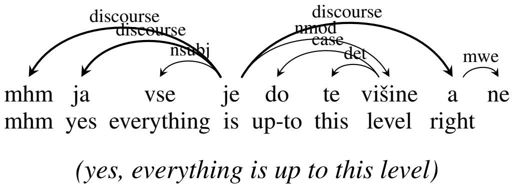

discourse
discourse
nsubj
nmod
discourse
case
det
mwe
mhm ja vse je do te višine a ne
mhm yes everything is up-to this level right
(yes, everything is up to this level)

Figure 4: Annotation of discourse elements by *discourse*.

The second group are connective adverbials that can appear in various syntactic positions without necessarily changing their discourse-relating meaning, such as *torej* 'thus', *zato* 'therefore', *vendar* 'however', *pa* (multi-functional particle), *se pravi* 'that is to say', *tako da* 'so', *s tem da* 'given that'. Following the principles of the JOS annotation scheme, we distinguish between different syntactic positions and label such connectives as coordinating conjunctions (*cc*) when used in clause- or sentence-initial position (*Končali smo, torej pojdimo domov.* 'We have finished, thus let's go home.') or as adverbial modifiers (*advmod*) when used in clause-medial positions (*Končali smo, pojdimo torej domov.* 'We have finished, let's thus go home.'). What is more, some of these items can also appear in utterance-initial position as markers of discourse organisation (*Torej, kako ste?* 'So, how are you?'), in which case we label them as *discourse*.

Similarly, expressions of modality, such as *seveda* 'of course', *v bistvu* 'in fact', *pač* 'well', *res* 'really', *vseeno* 'still', are annotated as adverbial or nominal modifiers, unless they appear in prosodically distinct utterance-initial or phrase-medial positions (analogue to comma punctuation), in which case they are also annotated as *discourse*. In future work on the treebank and the UD guidelines in general, however, we might reconsider the extent of differentiation between the various functions of discourse-related lexica on the morphological and syntactic layer, as well as the formal tests associated with it.

4.2.5. Sentential Parentheticals

Spoken utterances often include sentential parentheticals appearing in clause-medial position, and performing different appositional and commenting functions. These are labeled *parataxis* and attach to the main predicate, regardless of their (non-)projectivity. A separate *parataxis:discourse* label is introduced for the most frequent sentential parentheticals that have been grammaticalized into semantically bleached fixed expressions with discourse marking functions, such as *ne vem* 'I don't know', *(a) veš* 'you know', *mislim* 'I think', *recimo* 'say', *prosim* 'please', *glej* 'listen', and often appear in clause-medial positions, following the same attachment principles as non-clausal discourse markers.

Parenthetical clauses introduced by subordinating conjunctions, such as conditional speech acts (*če smem vprašati* 'if I may ask') or comment clauses (*če se prav spomnim*, 'if I remember correctly', *kolikor vem*, 'as far as I know', *kot rečeno* 'as previously said'), are labeled as adverbial clausal modifiers (*advcl*), regardless of their degree of grammaticalization.

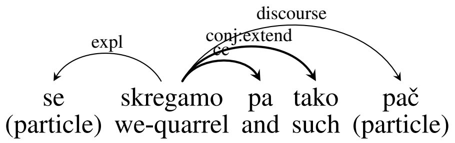
4.2.6. Asyndetic Coordination and General Extenders

Speakers often use strings of syntactically parallel constituents with no explicit coordinating conjunction between them, either as means of reformulation, specification or stylistic effect, or as a consequences of coordination being implied by prosody, e.g. *nobena muca, nobene miške, nobeni zajčki* 'no cats, no mice, no rabbits'. We annotate such asyndetic coordination structures as (*conj*), unless the parallelisms are instances of clausal juxtaposition or nominal apposition, in which the *parataxis* and *appos* labels are used, respectively.

Another group of frequent expressions treated as coordination are general extenders, expressions such as *in tako naprej* 'and so on', and *ali nekaj takega* 'or something like that', introduced by either coordinating or disjunctive conjunctions that typically attach to grammatically complete phrases or utterances. Although general extenders have a predominantly discourse managing function, we analyze them as a special type of coordinating conjuncts

[4] The various theoretical views on distinguishing and categorizing such items based on syntactic and/or semantic criteria also resonates in the current version of UD treebanks, with divergent part-of-speech tags and dependency labels across languages and inconsistent applications within treebanks.

(*conj:extender*), even though the heads of such pseudo-coordinations are sometimes syntactically asymmetrical (as in Figure 5). The same label is also used in annotation of tag questions introduced by disjunctive conjunctions (e.g. *ker sta ful grozna ali kaj?*, ’because they are so awful or what?’).

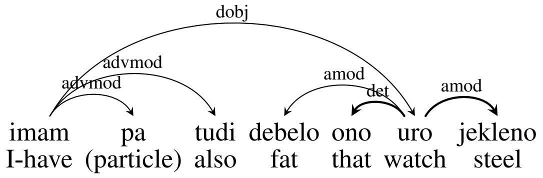


(we quarrel and such)

Figure 5: Annotation of general extenders by *conj:extend.*

4.2.7. Atypical Word Order

Due to posterior recall or the need for clarification, speakers ofter add individual syntactic units later than it would be expected by the word-order constraints of the written language, as in the example of postposed determiners and adjectival premodifier in Figure 6. With the exception of topic-marked fronted or postposed elements (labeled as *dislocated*), the dependency annotation remains independent of word-order particularities, regardless of potential non-projectivity.



(I also have that fat steel watch)

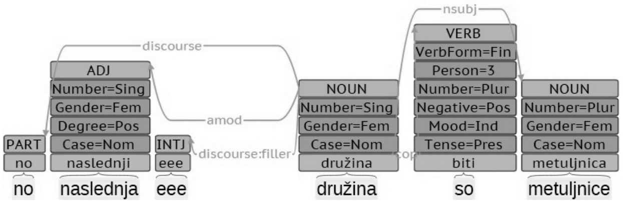
Figure 6: Atypical determiner and premodifier positions in a fixed-word-order NP.

5. Treebank Annotation

The annotation of the Spoken Slovenian Treebank was carried out in three consecutive steps. In the first step, lemmas and morphosyntactic tags were manually verified to correct mistakes by the statistical POS tagger and lemmatizer (Grčar et al., 2012) used in the annotation of the reference Gos corpus and based on the JOS annotation scheme (Erjavec et al., 2010). The aligned audio recordings were accessed through the reference corpus web concordancer. [5]

In the second step, the manually verified morphological information was automatically converted to UD POS tags and morphological feature-value pairs using the mapping script developed for the conversion of the ssj500k reference Slovenian treebank (Krek et al., 2015) to the written Slovenian UD Treebank. The written treebank was then also used to induce the initial Slovenian UD parsing model and to parse the spoken treebank with gold-standard UD morphology, using the MaltParser data-driven dependency parser (Nivre et al., 2007).

In the last stage of the annotation process, the automatically parsed treebank was imported to WebAnno (Yimam et al., 2013), a general purpose web-based linguistic annotation tool, for final manual corrections. In addition to correcting parser mistakes and implementing speech-specific enhancements of the annotation scheme presented in the section above, syntax-dependent POS tags that could not have been adequately converted due to lack of syntactic information in the original corpus, such as auxiliaries and determiners, were also manually corrected. Proper names anonimized in audio recordings and transcribed as non-linguistic tokens (e.g. *[name:personal]*) were also given the missing POS and feature information. An example of an annotated utterance in WebAnno is illustrated in Figure 7.



(so the next family are fabaceae)

Figure 7: An example of an annotated utterance in WebAnno showing the full set of lemma, part-of-speech, morphology and dependency annotation layers.

6. Treebank Analysis

This section presents the initial analysis of the distribution of dependency relations in the Spoken Slovenian UD Treebank as summarized in Figure 8. To better illustrate the particularities of spoken communication in relation to the characteristics of Slovenian language in general, the analysis is made in comparison with the written Slovenian UD Treebank.

In addition to discrepancies due to the newly introduced extensions, such as *conj:extend, discourse:filler, parataxis:discourse and parataxis:restart*, which have not yet been retroactively implemented in the written treebank (but presumed to have a low overall frequency), the observed syntactic differences between the two modalities can be broadly categorized into two general groups: differences due to particularities of speech production and transcription (text structure) and differences due to the nature of spoken communication (text contents).

Besides the divergence in the frequency of punctuation symbols (*punct*) due to specifics of Gos transcriptions that do not include punctuation characters, the most noticeable structural distinction is the difference in the size of syntactic trees (proportional number of *root* nodes), since utterances in the spoken treebank are typically shorter than sentences in the written treebank (with an average length of 9.2 and 17.7 tokens respectively). The high frequency of *parataxis* relation in the spoken treebank further suggests that spoken utterances are not only shorter but also more fragmented, containing sequences of juxtaposed sentences

[5] http://www.korpus-gos.net

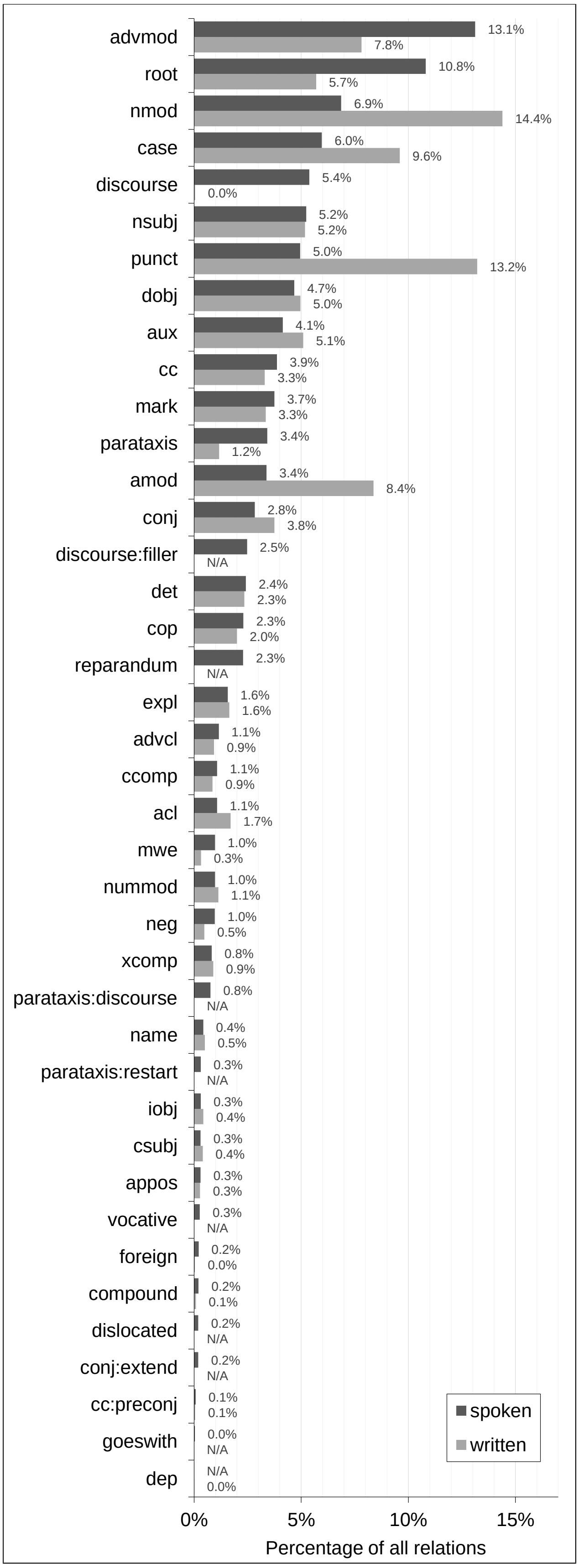


Figure 8: Comparison of dependency relations in the spoken and written Slovenian UD treebanks. N-spoken = 29,488[7]; N-written = 140,418.

without any explicit conjunction.[8] Furthermore, although the *root* node is much more frequent in spoken than in written language, *nsubj* and *dobj* have very similar frequencies. This indicates that, while spoken language has shorter sentences, many of them are either elliptic or do not contain a verbal predicate with a subject and an object.

Other structural particularities of the spoken treebank include high numbers of repairs (*reparandum*) and fillers (*disocurse:filler*), absent from the written treebank, as well as a higher frequency of dislocated sentence elements (*dislocated*) and spelled-out numbers (*compound*) in comparison with the written treebank.

As expected, content-dependent differences show a distinctively high number of elements of interaction, such as *discourse* and *vocative* relations, in comparison with the written treebank, where such constructions mainly appear in spoken-like dialogues. The comparison also shows that written Slovenian is significantly more nominal than spoken Slovenian with a higher share of nominal and prepositional phrases (the *nmod* and *case* relations) that are also more complex in terms of the number of adjectival (*amod*) or clausal (*acl*) attributives. On the other hand, the spoken treebank contains notably more adverbial modifiers (*advmod*), which mostly include expressions of deixis, discourse cohesion and modality. Grammaticalized expressions of interaction, discourse relations and modality are also those that constitute the majority of speech-frequent multi-word expressions (*mwe*).

7. Format and Availability

The first version of the Spoken Slovenian Treebank is planned to be released as part of the UD release v1.3 under the CC-BY-NC-SA 4.0 licence. In addition to the standard metalinguistic information specified by the CONLL-U format,[9] such as normalized word forms, lemmas, UD POS tags, JOS morphosyntactic tags, UD features and UD dependencies, the treebank also includes information on pronunciation-based word form transcriptions and tokenization (as part of the MISC column), thus ensuring compatibility with the original Gos corpus in TEI XML format. All metadata on individual utterances, such as information about text type, speaker demographics, recording region, channel etc., can be accessed through the unique utterance identifier in the comment line pointing to its full description in the original TEI header.

8. Conclusions and Future Work

The construction of the first dependency treebank of spoken Slovenian resulted in several important contributions

[8] As we have already explained in Section 4.2.4, the lack of explicit coordinating conjunctions does not, however, indicate the absence of cohesion between paratactical sentences in speech, as discourse relations in Slovenian are often expressed with clause-medial constructions, annotated as adverbial modifiers (*advmod*). Their host clauses are currently annotated as *parataxis*, rather than instances of asyndetic coordination (*conj*).

[8] The difference in token count between the original treebank after sampling and the final annotated treebank is due to changes in tokenization (splitting of fused word forms).

[9] http://universaldependencies.org/format.html

both in the field of Slovenian language resources and spoken language resources in general, namely the adaptation of the Universal Dependencies annotation scheme to particularities of spoken communication, the construction of the treebank, and its manual lemmatization, morphology and dependency annotation. The initial analysis of the Spoken Slovenian Treebank in comparison with the written Slovenian UD treebank confirmed significant linguistic differences between the two modalities, which not only reasserts the importance of development of speech-specific language resources, but also motivates several lines of future work.

From the local perspective, future work on the treebank should include its continual expansion, revisions of the annotation scheme, addition of new layers of linguistic annotation, comprehensive corpus-based analysis of spoken communication in Slovenian, and its comparisons with written language. From the global perspective, the proposed Universal Dependencies annotation scheme could be applied to spoken language treebanks for other languages, to further consolidate a standardised annotation of universal syntactic phenomena in speech and enable contrastive linguistic analyses. Last but not least, the Universal Dependencies Treebank of Spoken Slovenian represents an especially valuable resource for future experiments in data-driven speech processing, including novel explorations in cross-modal and cross-lingual spoken language dependency parsing.

9. Acknowledgements

The work presented in this paper has been partially supported by the Young Researcher Programme of the Slovenian Research Agency and Parseme ICT COST Action IC1207 STSM grant.

10. Bibliographical References

Charniak, E. and Johnson, M. (2001). Edit detection and parsing for transcribed speech. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Degand, L., Cornillie, B., and Pietrandrea, P. (2013). *Discourse Markers and Modal Particles : Categorization and Description.* Pragmatics & Beyond New Series. John Benjamins Publishing Company.

Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In Nicoletta Calzolari, et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Fischer, K. (2006). *Approaches to Discourse Particles*. Studies in pragmatics. Elsevier.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Grčar, M., Krek, S., and Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. [Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene]. In T. Erjavec et al., editors, *Proceedings of the 8th Language Technologies Conference*, volume C, pages 89–94, Ljubljana, Slovenia, October. IJS.

Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptácek, J., Toman, J., and Uresová, Z. (2008). PDTSL: An annotated resource for speech reconstruction. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, pages 93–96. IEEE.

Hinrichs, E. W., Bartels, J., Kawata, Y., Kordoni, V., and Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, Artificial Intelligence, pages 550–574. Springer Berlin Heidelberg.

Honnibal, M. and Johnson, M. (2014). Joint incremental disfluency detection and dependency parsing. *Transactions of the Association for Computational Linguistics*, 2(1):131–142.

Johannessen, J. B. and Jørgensen, F. (2006). Annotating and parsing spoken language. In Peter Juel Henrichsen et al., editors, *Treebanking for Discourse and Speech*, volume 32 of *Copenhagen Studies in Language*, pages 83–104. Samfundslitteratur Press, Copenhagen.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., and Tchobanov, A. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken french. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 295–301, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

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

Marneffe, M.-C. D., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal stanford dependencies: a cross-linguistic typology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Nivre, J. (2015). Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Rasooli, M. S. and Tetreault, J. (2013). Joint parsing and disfluency detection in linear time. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 124–129, Seattle, Washington, USA, October. Association for Computational Linguistics.

Shriberg, E. (1996). Disfluencies in switchboard. In *Proceedings of International Conference on Spoken Language Processing*, volume 96, pages 11–14.

van der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B., and Schuurman, I. (2002). Syntactic analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., and Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47(4):1031–1048.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 213–218, Marrakech, Morocco. European Language Resources Association.

11. Language Resource References

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., and Holz, N. (2015). Training corpus ssj500k 1.4. Slovenian language resource repository CLARIN.SI.

Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bosco, C., Bowman, S., Celano, G. G. A., Connor, M., de Marneffe, M.-C., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Erjavec, T., Farkas, R., Foster, J., Galbraith, D., Ginter, F., Goenaga, I., Gojenola, K., Goldberg, Y., Gonzales, B., Guillaume, B., Hajič, J., Haug, D., Ion, R., Irimia, E., Johannsen, A., Kanayama, H., Kanerva, J., Krek, S., Laippala, V., Lenci, A., Ljubešić, N., Lynn, T., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., Mori, S., Nurmi, H., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Prokopidis, P., Pyysalo, S., Ramasamy, L., Rosa, R., Saleh, S., Schuster, S., Seeker, W., Seraji, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Simov, K., Smith, A., Štěpánek, J., Suhr, A., Szántó, Z., Tanaka, T., Tsarfaty, R., Uematsu, S., Uria, L., Varga, V., Vincze, V., Žabokrtský, Z., Zeman, D., and Zhu, H. (2015). Universal dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., and Erjavec, T. (2013). Spoken corpus Gos 1.0. Slovenian language resource repository CLARIN.SI.