Universal Dependencies for Swedish Sign Language

Robert Östling, Carl Börstell, Moa Gärdenfors, Mats Wirén
Department of Linguistics
Stockholm University
{robert,calle,moa.gardenfors,mats.wiren}@ling.su.se

Abstract

We describe the first effort to annotate a signed language with syntactic dependency structure: the Swedish Sign Language portion of the Universal Dependencies treebanks. The visual modality presents some unique challenges in analysis and annotation, such as the possibility of both hands articulating separate signs simultaneously, which has implications for the concept of *projectivity* in dependency grammars. Our data is sourced from the Swedish Sign Language Corpus, and if used in conjunction these resources contain very richly annotated data: dependency structure and parts of speech, video recordings, signer metadata, and since the whole material is also translated into Swedish the corpus is also a parallel text.

1 Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016b) has produced a language-independent but extensible standard for morphological and syntactic annotation using a formalism based on dependency grammar. This standard has been used to create the Universal Dependencies treebanks (Nivre et al., 2016a), which in its latest release at the time of writing (version 1.4) contains 64 treebanks in 47 languages—one of which is Swedish Sign Language (SSL, ISO 639-3: SWL), the topic of this article.

There are very few sign languages for which there are corpora. Most of the available sign language corpora feature only simple sign segmentation and annotations, often also with some type of translation into a spoken language (either as written translations or as spoken voice-overs). Sign language corpora with more extensive syntactic annotation is limited to Australian Sign Language, which contains some basic syntactic segmentation and annotation (Johnston, 2014). Apart from this, smaller parts of the corpora of Finnish Sign Language (Jantunen et al., 2016) and Polish Sign Language (Rutkowski and Łozińska, 2016), have had some syntactic segmentation and analysis, and another such project is under way on British Sign Language.[1]

To the best of our knowledge, we present the first dependency annotation and parsing experiments with sign language data. This brings us one step closer to the goal of bridging the gap in availability between written, spoken and sign language natural language processing tools.

2 Universal Dependencies

The Universal Dependencies project aims to provide uniform morphological and syntactic (in the form of dependency trees) annotations across languages (Nivre et al., 2016b).[2] Built on a language-universal common core of 17 parts of speech and 40 dependency relations, there are also language-specific guidelines which interpret and when necessary extend those in the context of a given language.

3 Swedish Sign Language

Swedish Sign Language (SSL) is the main sign language of the Swedish Deaf community.[3] It is estimated to be used by at least 10,000 as one of their primary languages, and is the only sign language to be recognized in Swedish law, giving it a special status alongside the official minor-

[1] http://www.bslcorpusproject.org/projects/bsl-syntax-project/

[2] Note that our work predates version 2 of the UD guidelines, and is based on the first version.

[3] Capital D “Deaf” is generally used to refer to the language community as a cultural and linguistic group, rather than ‘deaf’ as a medical label.

Proceedings of the 21st Nordic Conference of Computational Linguistics, pages 303–308,
Gothenburg, Sweden, 23-24 May 2017. ©2017 Linköping University Electronic Press

ity languages (Ahlgren and Bergman, 2006; Parkvall, 2015). The history of SSL goes back at least 200 years, to the inauguration of the first Deaf school in Sweden, and has also influenced the two sign languages of Finland (i.e. Finnish Sign Language and Finland-Swedish Sign Language) with which SSL can be said to be related (Bergman and Engberg-Pedersen, 2010).

4 Data source

The SSL Corpus Project ran during the years 2009–2011 with the intention to establish the first systematically designed and publicly available corpus of SSL, resulting in the SSL Corpus (SSLC). Approximately 24 hours of video data of pairs of signers conversing was recorded, comprising 42 signers of different age, gender, and geographical background, spanning 300 individual video files (Mesch, 2012). The translation and annotation work is still on-going, with new releases being made available online as the work moves forward. The last official release of the SSLC includes just under 7 hours of video data (Mesch et al., 2012) along with annotation files containing 53,625 sign tokens across 6,197 sign types (Mesch, 2016).

The corpus is annotated using the ELAN software (Wittenburg et al., 2006), and the annotation files are distributed in the corresponding XML-based `.eaf` format. Each annotation file contains tiers on which annotations are aligned with the video file, both video and annotation tiers being visible in the ELAN interface (see Figure 1). The SSLC annotation files currently include tiers for sign glosses, and others for Swedish translations. Sign glosses are written word labels that represent signs with approximate meanings (e.g. PRO1 for a first person pronoun). The sign gloss annotation tiers are thus segmented for lexical items (i.e. individual signs), and come in pairs for each signer—each tier representing one of the signer's hands (one tier for the so-called *dominant hand*, and another for the *non-dominant hand*) (Mesch and Wallin, 2015).[4] Sign glosses also contain a part-of-speech (PoS) tag which have been derived from manually correcting the output of a semi-automatic method for cross-lingual PoS tagging (Östling et al., 2015). The translation tier is segmented into longer chunks, representing stretches of discourse that can be represented by an idiomatic Swedish translation. However, the translation segmentations do not represent clausal boundaries in either SSL or Swedish (Börstell et al., 2014). More recently, a portion of the SSLC was segmented into clausal units and annotated for basic syntactic roles (Börstell et al., 2016), which led to the current UD annotation work. Figure 1 shows the basic view of the SSLC videos and annotations in the ELAN software, with tiers for sign glosses and translations on the video timeline.

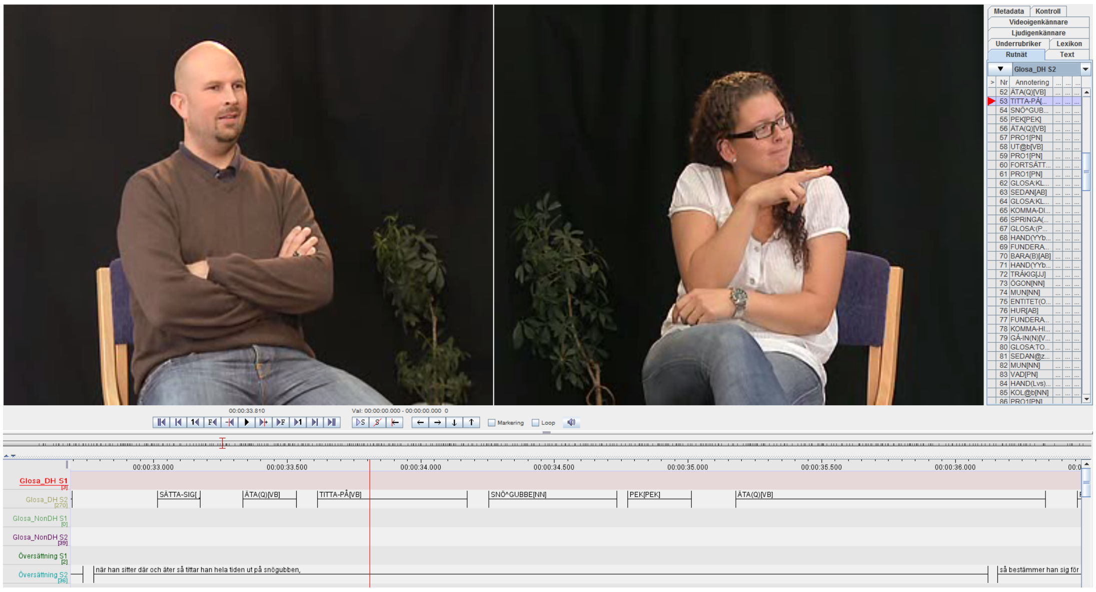
5 Annotation procedure and principles for SSL

For practical purposes, annotation was performed by extending the ELAN files of our source material from the SSLC project (see Figure 2 for an example). These annotations were automatically converted to the CoNLL-U format used by Universal Dependencies.

The annotation of UD based syntactic structure started by coming up with a procedure for annotating a signed language using ELAN. Signed language is more simultaneous than spoken language, particularly in the use of paired parallel articulators in form of the signer's two hands (Vermeerbergen et al., 2007). We handle this by allowing signs from both hands into the same tree structure, which leads to well-formed trees consistent with the dependency grammar formalism's single-head, connectedness and acyclicity constraints. These trees can however have some unusual properties compared to spoken languages. For the purpose of conforming to the CoNLL-U data format, which requires an ordered sequence of tokens, we sort signs by their chronological order. The chronological order spans both sign tiers per signer, and is defined as the onset time of each sign annotation. In the case of two signs on each hand tier (i.e. dominant vs. non-dominant hand) having identical onsets, favor is given to signs articulated by the signer's dominant hand. This working definition is by no means the only reasonable linearization, which means that the notion of projectivity to some extent loses its meaning. A tree can be considered projective or non-projective depending on how the ordering of simultaneously articulated signs is defined—assuming one wants to impose such an ordering in the first place.

Because the source material contains no segmentation above the sign level, we decided to use

[4] The dominant hand is defined as the hand preferred by a signer when signing a one-handed sign.



Nr	Annotering
52	ÄTA(Q)[VB]
53	TITTA-PÅ[...
54	SNÖ^GUB...
55	PEK[PEK]
56	ÄTA(Q)[VB]
57	PRO1[PN]
58	UT@b[VB]
59	PRO1[PN]
60	FORTSÄTT...
61	PRO1[PN]
62	GLOSA:KL...
63	SEDAN[AB]
64	GLOSA:KL...
65	KOMMA-DI...
66	SPRINGA(...
67	GLOSA:(P...
68	HAND(YYb...
69	FUNDERA...
70	BARA(B)[AB]
71	HAND(YYb...
72	TRÅKIG[JJ]
73	ÖGON[NN]
74	MUN[NN]
75	ENTITET(O...
76	HUR[AB]
77	FUNDERA...
78	KOMMA-HI...
79	GÅ-IN(N)[V...
80	GLOSA:TO...
81	SEDAN@z...
82	MUN[NN]
83	VAD[PN]
84	HAND(Lvs)...
85	KOL@b[NN]
86	PRO1[PN]

Glosa_DH S1 [3]

Glosa_DH S2 [270]: SÄTTA-SIG[| ÄTA(Q)[VB] | TITTA-PÅ[VB] | SNÖ^GUBBE[NN] | PEK[PEK] | ÄTA(Q)[VB]

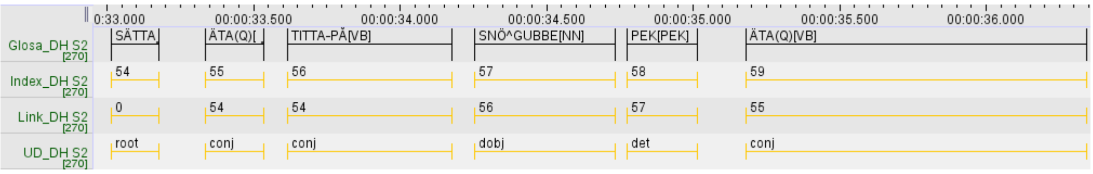
Glosa_NonDH S1 [0]

Glosa_NonDH S2 [39]

Översättning S1 [2]

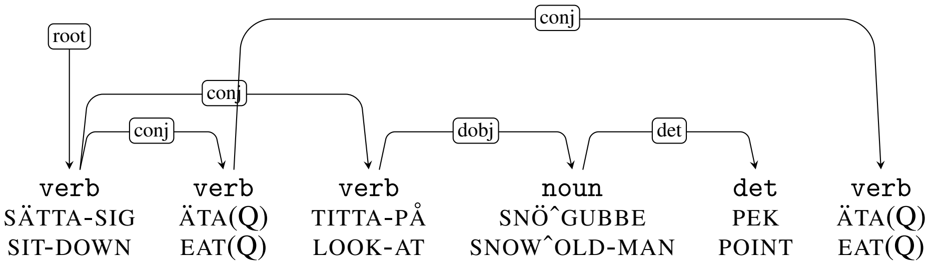
Översättning S2 [36]: när han sitter där och äter så tittar han hela tiden ut på snögubben, | så bestämmer han sig för

Figure 1: Screenshot of an SSLC file in ELAN. This is the material we base our dependency annotations on, and the annotator can easily view the source video recording.



	0:33.000	00:00:33.500		00:00:34.000	00:00:34.500	00:00:35.000	00:00:35.500
Glosa_DH S2 [270]	SÄTTA	ÄTA(Q)[	TITTA-PÅ[VB]	SNÖ^GUBBE[NN]	PEK[PEK]	ÄTA(Q)[VB]	
Index_DH S2 [270]	54	55	56	57	58	59	
Link_DH S2 [270]	0	54	54	56	57	55	
UD_DH S2 [270]	root	conj	conj	dobj	det	conj	

Figure 2: Screenshot zooming into the UD annotation tiers and sign–dependency linking for the utterance from Figure 1. This is the interface used by the annotator.



‘He is sitting there eating looking out at the snowman.’

Figure 3: The example from Figure 1 and Figure 2 with dependency annotations visualized. The (Q) suffix on the ÄTA(Q) gloss indicates which of the multiple signs for ‘eat’ in SSL is used in this case.

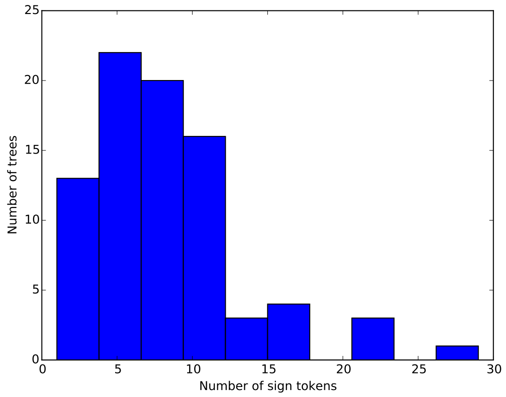


Figure 4: Distribution of tree sizes for the Swedish Sign Language Universal Dependencies treebank.

a bottom-up annotation procedure where subtrees were connected until we could find no more suitable mergers. In other words, the segmentation is entirely based on syntactic structure. The resulting fully connected trees were then used as “sentences” in the CoNLL-U format.

One peculiar feature of many sign languages is the repetition of verbs, sometimes referred to as *verb sandwiches*, in which one verb occurs in the middle of a sentence and also repeated at the end (Bergman and Dahl, 1994). Such a construction is found in Figure 3, in which the verb EAT appears in two places. Whereas verb chains (i.e. multiple verbs in one clause) were treated as coordinated elements linked to the `root` verb using the label `conj`, we decided to treat repeated verbs differently by labeling the repeated verb as a coordinated element linked to its first occurrence (see Figure 3).

6 Treebank statistics

The SSL treebank released in version 1.4 of the UD treebanks contains 82 trees with a total of 672 sign tokens (mean 8.2, median 7). The distribution of tree sizes (in tokens) is shown in Figure 4, as described in Section 5 these were produced in a bottom-up fashion and reflect our judgment of the largest sensible syntactic segmentation of the material. As could be expected from a corpus of spontaneous conversation, there is a large number of small trees. For comparison, the only spoken language (Slovenian) treebank has mean 9.1 and median 6, while the written Swedish treebank has mean 14.3 and median 13 sentence length, not counting punctuation.

7 Dependency parsing

Given that this is the first sign language UD treebank, we decided to perform some dependency parsing experiments to establish baseline results. We use the parser of Straka et al. (2015), part of the UDpipe toolkit (Straka et al., 2016), for our experiments. The training (334 tokens), development (48 tokens) and test (290 tokens) split from UD treebanks 1.4 was used. A hundred iterations of random hyperparameter search was performed for each of their parser models (projective, partially non-projective and fully non-projective), and the model with highest development set accuracy was chosen. Unsurprisingly given the small amount of training data, we found the most constrained projective model performed best, in spite of the data containing non-projective trees (see Figure 3). Development set attachment score was 60 and 56 (unlabeled and labeled, respectively) while the corresponding test set scores were 36 and 28. The discrepancy can be partly attributed to the much shorter mean sentence length of the development set: 6.0 vs 10.4 for the test set. Such low scores are not yet useful for practical tasks, but we emphasize that our primary goal in this work is to explore the possibility of UD annotation for a sign language. Our annotation project is ongoing, and we intend to further expand the SSL part in future UD treebanks releases.

8 Conclusions and future work

In releasing the Universal Dependencies treebank of Swedish Sign Language (SSL), the first such resource for a signed language, we hope to enable new computational research into sign language syntax. We have shown that even though some theoretical and practical issues exist when applying UD principles to a sign language, it is possible to come up with a reasonable annotation scheme. In the long run, we hope this will stimulate the development of Natural Language Processing (NLP) tools capable of processing sign languages. Finally, because we have both parallel data in Swedish and language-independent syntactic annotations, we also believe this resource could prove particularly useful in cross-lingual NLP.

Acknowledgments

Part of this work has been supported by an infrastructure grant from the Swedish Research Council (SWE-CLARIN, project 821-2013-2003).

References

Inger Ahlgren and Brita Bergman. 2006. Det svenska teckenspråket. In *Teckenspråk och teckenspråkiga: Kunskaps- och forskningsöversikt (SOU 2006:29)*, pages 11–70. Statens offentliga utredningar.

Brita Bergman and Östen Dahl. 1994. Ideophones in sign language? The place of reduplication in the tense-aspect system of Swedish Sign Language. In Carl Bache, Hans Basbøll, and Carl-Erik Lindberg, editors, *Tense, Aspect and Action. Empirical and Theoretical Contributions to Language Typology*, pages 397–422. Mouton de Gruyter.

Brita Bergman and Elisabeth Engberg-Pedersen. 2010. Transmission of sign languages in the Nordic countries. In Diane Brentari, editor, *Sign languages: A Cambridge language survey*, chapter 4, pages 74–94. Cambridge University Press, New York, NY.

Carl Börstell, Johanna Mesch, and Lars Wallin. 2014. Segmenting the Swedish Sign Language Corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, Eleni Efthimiou, Evita Fotinea, Thomas Hanke, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, pages 7–10, Paris. European Language Resources Association (ELRA).

Carl Börstell, Mats Wirén, Johanna Mesch, and Moa Gärdenfors. 2016. Towards an annotation of syntactic structure in Swedish Sign Language. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 19–24, Paris. European Language Resources Association (ELRA).

Tommi Jantunen, Outi Pippuri, Tuija Wainio, Anna Puupponen, and Jorma Laaksonen. 2016. Annotated video corpus of FinSL with Kinect and computer-vision data. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 93–100, Paris. European Language Resources Association (ELRA).

Trevor Johnston. 2014. The reluctant oracle: Adding value to, and extracting of value from, a signed language corpus through strategic annotations. *Corpora*, 9(2):155–189.

Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1):103–121.

Johanna Mesch, Lars Wallin, Anna-Lena Nilsson, and Brita Bergman. 2012. Dataset. Swedish Sign Language Corpus project 2009–2011 (version 1).

Johanna Mesch. 2012. Swedish Sign Language Corpus. *Deaf Studies Digital Journal*, 3.

Johanna Mesch. 2016. Annotated files for the Swedish Sign Language Corpus. Version 4.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2016a. Universal dependencies 1.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016b. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Robert Östling, Carl Börstell, and Lars Wallin. 2015. Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In Beata Megyesi, editor, *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015), NEALT Proceedings Series 23*, pages 263–268, Vilnius. ACL Anthology.

Mikael Parkvall. 2015. *Sveriges språk i siffror: vilka språk talas och av hur många?* Språkrådet.

Paweł Rutkowski and Sylwia Łozińska. 2016. Argument linearization in a three-dimensional grammar: A typological perspective on word order in Polish Sign Language (PJM). *Journal of Universal Language*, 17(1):109–134.

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič jr. 2015. Parsing universal dependency treebanks using neural networks and search-based oracle. In *Proceedings of Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*, December.

Milan Straka, Jan Hajič, and Straková Jana. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France, May. European Language Resources Association (ELRA).

Myriam Vermeerbergen, Lorraine Leeson, and Onno Crasborn, editors. 2007. *Simultaneity in signed languages: Form and function*. John Benjamins, Amsterdam/Philadelphia, PA.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.