The LIA Treebank of Spoken Norwegian Dialects

Lilja Øvrelid[†], **Andre Kåsen** [†‡], **Kristin Hagen**[‡], **Anders Nøklestad**[‡],
Per Erik Solberg[*] **and Janne Bondi Johannessen**[‡]
[†] Department of Informatics, University of Oslo
[‡] Department of Linguistics and Scandinavian Studies, University of Oslo
[*] Department of Philosophy, Classics, History of Arts and Ideas, University of Oslo
{liljao,andrekaa}@ifi.uio.no, {kristin.hagen,j.b.johannessen, anders.noklestad}@iln.uio.no, p.e.solberg@ifikk.uio.no

Abstract

This article presents the LIA treebank of transcribed spoken Norwegian dialects. It consists of dialect recordings made in the period between 1950–1990, which have been digitised, transcribed, and subsequently annotated with morphological and dependency-style syntactic analysis as part of the LIA (Language Infrastructure made Accessible) project at the University of Oslo. In this article, we describe the LIA material of dialect recordings and its transcription, transliteration and further morphosyntactic annotation. We focus in particular on the extension of the native NDT annotation scheme to spoken language phenomena, such as pauses and various types of disfluencies, and present the subsequent conversion of the treebank to the Universal Dependencies scheme. The treebank currently consists of 13,608 tokens, distributed over 1396 segments taken from three different dialects of spoken Norwegian. The LIA treebank annotation is an on-going effort and future releases will extend on the current data set.

Keywords: treebanks, spoken language, dialects, Norwegian, Universal Dependencies

1. Introduction

Large-scale initiatives like the CoNLL shared tasks on dependency parsing (Surdeanu et al., 2008), the Universal Dependencies (UD) initiative (Nivre et al., 2016) and the recent shared task on multilingual parsing from raw text (Zeman et al., 2017) have made available syntactic treebanks for a large number of languages, thus enabling parsing research for a wide variety of languages. Available treebanks are still, however, largely based on written textual resources, with a few exceptions (Dobrovoljc and Nivre, 2016; Östling et al., 2017).

The LIA project[1] has as its main objective to create a corpus consisting of old dialect recordings and make these accessible for research in linguistics and digital humanities. By digitization, transcription and further linguistic processing, this corpus can play an important role in the diachronic study of Norwegian dialects and more generally the linguistic variation in Norway. This article describes the LIA treebank of spoken Norwegian dialects. A longterm goal of this work is to develop a parser for spoken Norwegian, with the immediate goal of parsing the whole LIA material. This will enable more fine-grained linguistic analyses to be carried out over the material.

In this paper we present the LIA data set, its transcription and subsequent morphological and syntactic annotation, with a focus on the extended annotation guidelines of the Norwegian Dependency Treebank (NDT) for spoken language phenomena and the conversion of the treebanked data to the Universal Dependencies (UD) scheme (de Marneffe et al., 2014). The UD version of the treebank was made available with the v2.1 release of the UD treebanks (Nivre et al., 2017).

2. The LIA material

The LIA project (Language Infrastructure made Accessible) is a five-year national collaboration project between four Norwegian universities (University of Oslo, University of Bergen, University of Tromsø and The Norwegian University of Science and Technology), Norsk ordbok 2014 and Språkbanken at the National Library, in addition to international partners.

The main aim of the LIA project is to collect dialect recordings from the four participating universities, digitise them, inventorise, catalogue and safely store them and make them accessible for further research. The most interesting recordings are transcribed and text-sound synchronised with the transcription tool ELAN[2]. Finally they are morphologically tagged and parsed. This process is described below. The final outcome of the LIA project is a user-friendly searchable dialect corpus.

The audio files that constitute the data set are recorded between 1950 and 1990 in order to explore and survey the many different dialects in Norway. Sometimes the research questions also concern person or place names. Most of the informants are older people who are native speakers of their dialect. Typically, the recordings are interviews about old trades such as agriculture, fisheries, logging and life at the summer farm. Other topics are weaving, knitting, baking or dialects. The recordings are semi-formal or informal and often take place in an informant's home.

2.1. Transcription

The LIA project makes use of a semi-phonetic transcription standard, similar to that of Papazian and Helleland (2005) and described in Hagen et al. (2017). This standard is chosen mainly to conserve particularities in the different dialects.

The speech flow is separated into what we call *segments*. A segment is our spoken language approximation of a sen-

[1] http://www.hf.uio.no/iln/english/research/projects/language-infrastructure-made-accessible

[2] https://tla.mpi.nl/tools/tla-tools/elan

tence. Few special characters are in use. The exceptions are the Norwegian letters æ, ø and å, quotation marks for indicating indirect speech, '#' signifying a pause in the speech flow and variations of '+' and '%' combined with a letter indicating unclear speech or laughter etc. The '%' character followed by a letter represents an independent incident in the speech flow like laughter, coughing etc. The '+' characterizes the following word or word group. '+u' means for instance that the following word(s) are unclear. '+x' means that the word(s) are not listed in the dictionary. The variants of '+' and '%' are stripped from the transcripts prior to further morphosyntactic processing. This is done under the assumption that these phenomena do not have any syntactic significance. They are inserted back into the transcripts before the transcripts are made available for search online.

2.2. Transliteration

Before tokenization and lemmatization the semi-phonetic transcriptions are semi-automatically transliterated to standard Norwegian Nynorsk[3] orthography by the Oslo Transliterator[4]. The transliterator can be trained to transliterate any dialect or language variety into any other orthographical representation, and it is so far trained on more than 100 Norwegian dialects in the LIA project. The outcome from the transliterator is manually corrected and the resulting pair of transcriptions are used for training the transliterator for this particular dialect, improving performance on subsequent transliterations of that dialect.

The Oslo Transliterator has a web interface where the transcriptions can be uploaded and associated with the appropriate dialect. The transcriptions are divided into smaller parts, which are transliterated one by one. Each part is manually corrected and added to the training material before the transliterator is trained once more and performs better on the next transcription part. The results of each iteration of the training process are stored in a MySQL database. When all parts are completed, the transcriptions can be downloaded as ELAN files with the semi-phonetic transcription and the orthograpic transcription as separate layers.

New dialects can also be registered in the web interface. Instead of starting from scratch on the new dialect, the transliterator employs a technique in which suggestions for transliterated word forms for the new dialect are based on combinations of stored word form correspondences from a set of dialects selected among those that are already transliterated. Each of those dialects is given a weight based on how similar it is perceived to be to the new dialect by the human transliterator. This technique enables us to take advantage of various degrees of dialect similarity without requiring large amounts of training data or labour-intensive work on creating string similarity mappings.

Head	Dependent
Preposition	Prepositional complement
Finite verb	Complementizer
First conjunct	Subsequent conjuncts
Finite auxiliary	Lexical/main verb
Noun	Determiner

Table 1: Annotation choices in the NDT

3. Morphosyntactic annotation

For grammatical phenomena which are not specific to spoken language, we have followed the annotation scheme of the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014). An important reason for this choice was the detailed language-specific annotation guidelines which were developed for the NDT project (Kinn et al., 2014). These guidelines are custom-made for Norwegian, following the Norwegian Reference Grammar (Faarlund et al., 1997) closely. Furthermore, the NDT scheme has performed well in previous measures of inter-annotator agreement. Solberg et al. (2014) report agreement scores of 96.8% unlabeled and 95.3% labeled accuracy and Skjærholt (2014) quantified inter-annotator agreement using a chance-corrected metric derived from Krippendorff's α and showed that agreement on the NDT data is high: scoring an α of about 98%, among the highest of all the data sets studied. This annotation scheme was therefore a natural choice for the current project. An automatic conversion procedure to Universal Dependencies has furthermore been developed for the written NDT data set (Øvrelid and Hohle, 2016; Velldal et al., 2017). It is in other words possible to convert the LIA treebank to the UD annotation scheme, with a few modifications, see section 5.

3.1. The NDT scheme

The Norwegian Dependency Treebank contains manually annotated syntactic and morphological information for both varieties of Norwegian. The part-of-speech annotation follows the Oslo-Bergen Tagger scheme (Hagen et al., 2000). This scheme also marks inflectional features such as tense, number, gender and categories such as demonstrative and quantifier. As mentioned above, the syntactic annotation scheme is, to a large extent, based on the Norwegian Reference Grammar (Faarlund et al., 1997) and the dependency representations are inspired by choices made in comparable treebanks, in particular the Swedish treebank Talbanken (Nivre et al., 2006). Table 1 summarizes the main annotation choices concerning head status and dependency graphs in NDT.

3.2. Preprocessing

The transliterated transcripts are tokenized with whitespace as token delimiter and time code as segment delimiter. Quotation marks and '#'s are considered to be tokens. Lemmatization is completed with lemmas from Norsk ordbank, a lexicographic database for Norwegian.

[3] Norwegian has two official orthographic standards: Bokmål and Nynorsk. For the LIA transcriptions we have chosen to use Nynorsk, the standard closest to most Norwegian dialects.

[4] http://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/oslo-transliterator/

	Segments	Tokens
Eidsberg	679	5880
Vardø	450	5361
Austevoll	267	2367
total	1396	13608

Table 2: Raw counts for the different informants in the data set.

3.3. Annotation process

Prior to the manual morphosyntactic annotation, the LIA data set was automatically tagged with OBT+stat, a rule-based Constraint Grammar tagger with a HMM-based overlay (Johannessen et al., 2012) and parsed with the MATE parser (Bohnet, 2010) trained on the Nynorsk part of NDT, which consists largely of newspaper text. This parser has been reported to achieve a labeled accuracy score (LAS) of 89.54 on the Nynorsk test set of NDT (Solberg et al., 2014). The automatic tag assignments are then corrected by trained linguists using a browser-based application[5]. The dependency analyses are also manually corrected, following the extended guidelines described in section 4. below. Dependency annotation was performed using the TrEd application, which is the annotation tool developed for the annotation of the Prague Dependency Treebank (Böhmová et al., 2003).

3.4. Treebank data

Our data set, at present, consists of elderly speakers (80+) of the Eidsberg, Austevoll and Vardø dialects. This represents a diverse set of dialects from different regions of the country. Table 2 presents the number of segments and tokens, and their distributions across the three different dialects.

4. The LIA annotation guidelines

Spoken language contains several phenomena that distinguish it from written language, such as various types of disfluencies, repetitions and deletions (Shriberg, 1996; Johannessen and Jørgensen, 2006). Spoken language furthermore contains a larger number of fragmentary segments than written text. In the LIA guidelines we extend the annotation scheme of NDT with dependency analyses of syntactic phenomena that are specific to spoken language. In the following we will describe the main additions to the NDT scheme. These are further summarized in table 3, which shows the added part-of-speech tags and dependency relations. Figure 1 shows the dependency graph for an example sentence from the treebank.

4.1. Spoken language PoS

In order to account for spoken language, some additional PoS tags are added to the tagset. Incomplete or interrupted words are tagged with the tag `ufullst`, pauses ('#') with the tag `pause` and filled pauses or hesitations with `nol`. The category of interjections, we found, is quite frequent in our material. Therefore, a list of standardized interjections has been compiled and these receive the existing NDT tag for interjections (`interj`), see Figure 1 which includes the interjection *å*. Another issue is the preproprial article (investigated in Håberg (2010)) which is wide-spread in spoken Norwegian, but not common in written text. These are assigned the pronominal part-of-speech `pron` but function syntactically as a determiner (`DET`).

4.2. Spoken language syntax

The extended annotation guidelines for the syntactic annotation of the LIA material is built on the work of Dobrovoljc and Nivre (2016), who describe the annotation of the Slovenian spoken language treebank with Universal Dependencies. Below we describe our treatment of extra-linguistic tokens, various types of disfluencies, ellipsis and discourse elements.

4.2.1. Extra-linguistic tokens

During transcription, some extra-linguistic tokens are introduced in order to mark phenomena such as pauses or unfinished/incomplete words. The examples in (1)-(3) illustrate phenomena that introduce extra-linguistic tokens and will be discussed further below.

(1) *ja # og køyrde mjølka ut i byen igjen*
yes # and drove milk out in town again
'Yes # and drove the milk to town again'

(2) *å det var e det var noko*
oh it was mm it was something
forferdeleg trafikk
terrible traffic
'Oh there was mm there was terrible traffic'

(3) *så det var mykje g- e mykje greier*
so it was much t- mm much things
'So there were many t- mm many things'

Pauses (#), as in (1), and filled pauses (*e* 'mm'), as in (2), are treated similarly in the dependency structure, and are attached to the following dependent. In cases where there is more than one possible attachment site, we attach as high in the tree as possible, while keeping with a projective analysis. These extra-linguistic tokens are assigned the filler dependency relation (`FYLL`), see figure 1. Incomplete or interrupted words, as in example (3), are marked during transcription with a hyphen word-finally. If the incomplete word bears no relation to the surrounding context, it is given the syntactic function `FYLL`, otherwise it is treated as part of a repair relation, see section 4.2.2.

4.2.2. Disfluencies

We distinguish two types of disfluencies in our annotation of the LIA material: repairs and deletions, and introduce two new dependency relations (`REP` and `SLETT`) to account for these.

A repair consists of two parts: the reparandum and the repair. The reparandum is attached to its repair, which is always to the right of it. The repair relation `REP` is used both for repetitions, substitutions and reformulations. The reparandum will have the `REP`-relation to its repair as in

[5] `http://github.com/andrely/tag-annotator`

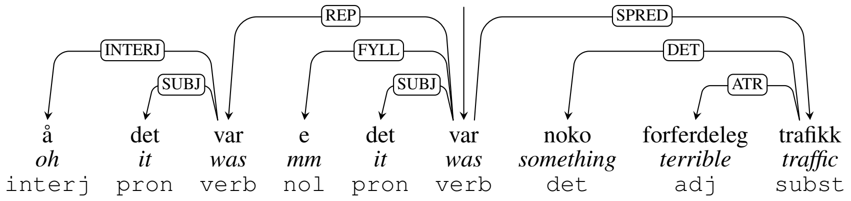

Figure 1: Example sentence from the LIA treebank with corresponding English gloss, PoS and dependency analysis.

	NDT Addition	Description	UD conversion
PoS	`nol`	filled pauses	`X`
	`ufullst`	incomplete words	`X`
	`pause`	pauses	`PUNCT`
Deprel	`FYLL`	fillers	`discourse:filler`
	`REP`	repairs	`reparandum`
	`SLETT`	deletions	`parataxis:deletion`

Table 3: Overview of additions to the NDT schemes in terms of part-of-speech tags and dependency relations in the LIA treebank, along with their converted UD relation.

figure 1, which shows the dependency graph for the example sentence in (2), where *det var* 'it was' is repeated. Note that a repair relation will only be used if there is some shared content between the reparandum and the repair. In the example in figure 1, we see an example of a repetition, where the repair repeats part of the reparandum. Otherwise, the deletion (`SLETT`) relation should be employed.

A deletion is distinguished from a repair by being semantically unrelated to the subsequent material. Example (4) illustrates a deletion, where the initial part of the sentence *måtte du* 'did you have to' is followed directly by the unrelated sentence *det var rasjonert* 'it was rationed'.

(4) *måtte du it var rasjonert?*
must you it was rationed?
'Did you have to it was rationed?'

Both deletions and repairs are attached as high as possible in the ensuing structure with which it is related, preserving projectivity. Our treatment of deletions departs from that of Dobrovoljc and Nivre (2016), who denote these as "restarts" and choose the incomplete element as head of the ensuing structure (the restart). We follow Shriberg (1996) in naming these deletions and we attach the deleted segment to the restart (the ensuing complete part of the utterance), which is situated to the right. So for example (4) above, the verb *måtte* 'must' is attached to the following finite verb *var* 'was'. This is thus similar to the analysis of discourse fillers (filled pauses), see below, and we preserve the overall structure of the segment despite the initial or internal incomplete structure.

4.2.3. Ellipsis

Ellipsis is a quite common phenomenon in spoken material (Johannessen and Jørgensen, 2006). The LIA treebank follows the treatment of ellipsis adopted in the NDT treebank. It does not introduce empty nodes. So, if the subject of a clause is ellided, as in example (5), there is simply no subject dependent.

(5) *her i bygda var # tretten gardsbruk ##*
here in town was # thirteen farms ##
og to #plassmenn
and two #place-men
'Here in town were thirteen farms and two smallholders'

Fragmentary segments are also common in spoken language (Shriberg, 1996). Segments that lack a finite verb are analyzed as fragments, using the `FRAG`-relation from the NDT scheme. We follow the prominence hierarchy provided in Kinn et al. (2014) in order to determine the head of the segment. It states that in the absence of a finite verb, head status should be given to non-finite verbs. If there is no non-finite verb, the most prominent element is the subject, followed by indirect objects or subject predicatives, etc. The same hierarchy is employed for cases of verbal ellipsis in coordination, where we follow the NDT guidelines in assigning a dedicated dependency relation `KOORD-ELL` to the remaining argument.

4.2.4. Discourse elements

For the treatment of interjections, we follow the NDT guidelines, which assign the dependency relation `INTERJ` to these elements, see figure 1. Interjections may often also constitute the root of a segment, a phenomenon which is common in spoken language.

Discourse fillers (or filled pauses in the terminology of Shriberg (1996)) are assigned a separate part-of-speech tag `nol` and given the dependency relation `FYLL`. These elements are attached to the right, as illustrated by the dependency graph in Figure 1, where the discourse filler *e* 'mm' is attached to the following finite verb *var* 'was' with the `FYLL` relation.

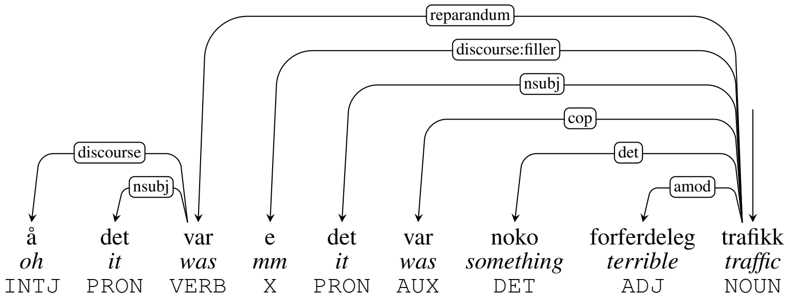

Figure 2: Example sentence from the UD conversion of the LIA treebank with corresponding English gloss, PoS and dependency analysis.

5. Conversion to Universal Dependencies

Universal Dependencies builds on several previous initiatives for universally common morphological (Zeman, 2008; Petrov et al., 2012) and syntactic dependency (McDonald et al., 2013; Rosa et al., 2014) annotation. Among its main tenets is the primacy of content words, i.e., content words, as opposed to function words, are syntactic heads wherever possible. It is intended to be a universal annotation scheme, i.e., applicable to any language, however it also offers some possibilities for language-specific information. With reference to the NDT annotation choices in Table 1, the UD scheme adopts the reverse attachment for auxiliaries, infinitival markers and prepositions.

The NDT and UD schemes differ in terms of both PoS tagset and morphological features, as well as structural analyses. The conversion therefore requires non-trivial transformations of the dependency trees, in addition to mappings of tags and labels that make reference to a combination of various kinds of linguistic information. For instance, in terms of PoS tags, the UD scheme offers a dedicated tag for proper nouns (PROPN), whereas NDT expresses information about noun type among its morphological features. UD further distinguishes auxiliary verbs (AUX) from main verbs (VERB). This distinction is not explicitly made in NDT, hence the conversion procedure makes use of the syntactic context of a verb; verbs that have a non-finite dependent are marked as auxiliaries. Further details about the conversion is given in Øvrelid and Hohle (2016), as well as in Velldal et al. (2017), which describes the extension of the conversion to cover the Nynorsk variant of Norwegian.

When it comes to part-of-speech tags, the universal tagset must be employed and there are few possibilities for language-specific adaptation. For dependency relations, there is the possibility to add treebank-specific subtypes of the universal dependency relations (on the form `udep:subtype`). Table 3 shows the treatment of the spoken language specific PoS tags and dependency relations during conversion to UD. Hesitations, as in example (2), and incomplete words, as in (3), are assigned the PoS tag `X` which is used for unknown words in UD and is the tag chosen by Dobrovoljc and Nivre (2016) for these phenomena. Pauses, marked by #, are assigned the PoS tag PUNCT.

For the conversion of the `FYLL` relation, we follow Dobrovoljc and Nivre (2016) in mapping directly to the universal relation `discourse`, with the subtype `filler`. Repairs are also straight-forwardly converted to the UD relation `reparandum`. For the analysis of restarts, or deletions as we have called them, we introduce a subtype of the universal `parataxis` relation called `deletion`. Figure 2 shows the converted UD version of the sentence from Figure 1. We observe that the structure differs markedly from the structure in the NDT format. The NDT version in Figure 1 annotates the finite verb *var* 'was' as the root of the segment, whereas the UD version appoints the predicative argument *trafikk* 'traffic' as root with the verb as a dependent with the `cop` (copula) relation type.

6. Availability of the treebank

The treebank will be made available for searching in Glossa (Nøklestad et al., 2017), which is a web-based corpus search interface being developed at the Text Laboratory, University of Oslo. This interface, which currently only supports searching in morphosyntactic information, will be extended with capabilities for searching in syntactic dependency structures as well.

For syntactic search we aim to implement an example-based approach along the lines of the GrETEL system[6], where the user can input an example of the kind of construction they are interested in, have the system analyse the example, select the relevant parts of the analysis (e.g. particular syntactic or morphosyntactic categories, lemmas and/or concrete word forms), and receive a list of all constructions in the treebank that match the given search criteria.

The Universal Dependencies version of the data set has been made available with the v2.1 release of the UD treebanks (Nivre et al., 2017). The treebank annotation continues and future releases will extend the treebank presented in this article with more data from more dialects.

7. Conclusion

In this article we have introduced the LIA treebank of spoken Norwegian dialects. The treebank currently consists of 13,608 tokens taken from three different dialects. We have presented our extended guidelines for morphological and syntactic annotation, as well as the conversion of the treebank to the Universal Dependencies scheme.

[6] `http://gretel.ccl.kuleuven.be/gretel3/`

8. Bibliographical References

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank. In *Treebanks: Building and Using Parsed Corpora*, pages 103–127. Kluwer, Dordrecht, the Netherlands.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the International Conference on Computational Linguistics COLING*, pages 89–97, Beijing, China.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. D. (2014). Universal Stanford dependencies. A cross-linguistic typology. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 4585–4592, Reykjavik, Iceland.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1566–1573, Portorož, Slovenia.

Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Universitetsforlaget, Oslo.

Håberg, L. (2010). Den preproprielle artikkelen i norsk.

Hagen, K., Johannessen, J. B., and Nøklestad, A. (2000). A constraint-based tagger for Norwegian. In *17th Scandinavian Conference in Linguistics*, pages 31–48.

Hagen, K., Håberg, L., Olsen, E., and Søfteland, Å. (2017). Transkripsjonsrettleiing for LIA. Technical report, The Text Laboratory, University of Oslo.

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

Johannessen, J. B., Hagen, K., Lynum, A., and Nøklestad, A., (2012). *OBT+stat: A combined rule-based and statistical tagger*, volume 49, page 51. John Benjamins, Amsterdam, the Netherlands.

Kinn, K., Solberg, P. E., and Eriksen, P. K. (2014). NDT guidelines for morphological and syntactic annotation. Technical report, National Library of Norway, Oslo.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., and Täckström, O. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of Association for Computational Linguistics (ACL)*, pages 92–97.

Nivre, J., Nilsson, J., and Hall, J. (2006). Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

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 International Conference on Language Resources and Evaluation (LREC)*.

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Bauer, J., Bengoetxea, K., Bhat, R. A., Bick, E., Bobicev, V., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Burchardt, A., Candito, M., Caron, G., Cebiroğlu Eryiğit, G., Celano, G. G. A., Cetin, S., Chalub, F., Choi, J., Cinková, S., Çöltekin, Ç., Connor, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eli, M., Elkahky, A., Erjavec, T., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grioni, M., Grūzītis, N., Guillaume, B., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Harris, K., Haug, D., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Ion, R., Irimia, E., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kanayama, H., Kanerva, J., Kayadelen, T., Kettnerová, V., Kirchner, J., Kotsyba, N., Krek, S., Laippala, V., Lambertino, L., Lando, T., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., Mendonça, G., Miekka, N., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, S., Moskalevskyi, B., Muischnek, K., Müürisep, K., Nainwani, P., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikolaev, V., Nurmi, H., Ojala, S., Osenova, P., Östling, R., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Perrier, G., Petrov, S., Piitulainen, J., Pitler, E., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Rademaker, A., Ramasamy, L., Rama, T., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rinaldi, L., Rituma, L., Romanenko, M., Rosa, R., Rovati, D., Sagot, B., Saleh, S., Samardžić, T., Sanguinetti, M., Saulīte, B., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Szántó, Z., Taji, D., Tanaka, T., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Washington, J. N., Wirén, M., Wong, T.-s., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., and Zhu, H. (2017). Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nøklestad, A., Hagen, K., Johannessen, J. B., Kosek, M., and Priestley, J. (2017). A modernised version of the glossa corpus search system. In Jörg Tiedemann, editor, *Proceedings of the 21st Nordic Conference on Computa-*

tional Linguistics, pages 251–254.

Östling, R., Börstell, C., Gärdenfors, M., and Wirén, M. (2017). Universal dependencies for swedish sign language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 303–308.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Papazian, E. and Helleland, B. (2005). *Norsk talemål. Lokal og sosial variasjon.* Høyskoleforlaget, Oslo.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2089–2096.

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtský, Z. (2014). Hamledt 2.0: Thirty dependency treebanks stanfordized. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341.

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

Skjærholt, A. (2014). A chance-corrected measure of inter-annotator agreement for syntax. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannesen, J. B. (2014). The Norwegian Dependency Treebank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL 2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. In *Proceedings of the Conference on Natural Language Learning*.

Velldal, E., Øvrelid, L., and Hohle, P. (2017). Joint ud parsing of norwegian bokmål and nynorsk. In Jörg Tiedemann, editor, *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10.

Zeman, D., Popel, M., Straka, M., Hajic, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinkova, S., Hajic jr., J., Hlavacova, J., Kettnerová, V., Uresova, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C. D., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., dePaiva, V., Droganova, K., Martínez Alonso, H., Çöltekin, c., Sulubacak, U., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Alcalde, H. F., Strnadová, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonca, G., Lando, T., Nitisaroj, R., and Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.