The INESS Treebanking Infrastructure

Paul Meurer[2], *Helge Dyvik*[1,2], *Victoria Rosén*[1,2], *Koenraad De Smedt*[1], *Gunn Inger Lyse*[1], *Gyri Smørdal Losnegaard*[1], *Martha Thunes*[1]

(1) University of Bergen, Norway
(2) Uni Computing, Bergen, Norway

paul.meurer@uni.no, helge.dyvik@uib.no, victoria@uib.no, desmedt@uib.no, gunn.lyse@lle.uib.no, gyri.losnegaard@lle.uib.no, martha.thunes@lle.uib.no

ABSTRACT

This paper briefly describes the current state of the evolving INESS infrastructure in Norway which is developing treebanks as well as making treebanks more accessible to the R&D community. Recent work includes the hosting of more treebanks, including parallel treebanks, and increasing the number of parsed and disambiguated sentences in the Norwegian LFG treebank. Other recent improvements include the presentation of metadata and license handling for restricted treebanks. The infrastructure is fully operational and accessible, but will be further improved during the lifetime of the INESS project.

KEYWORDS: treebanks, research infrastructure, parsed corpora, metadata, IPR, INESS, METANORD, CLARIN, CLARINO.

1 Introduction

This short paper sketches the current state of the *Infrastructure for the Exploration of Syntax and Semantics* (INESS) (Rosén et al., 2012a). The implementation and operation of this infrastructure is carried out by the University of Bergen (Norway) and Uni Computing (a division of Uni Research, Bergen), and is funded by the Research Council of Norway and the University of Bergen (2010–2016).

INESS is aimed at providing access to treebanks to the R&D community in the language sciences. One of the project's main activities is the development of a large, deep parsebank for Norwegian with a wide coverage grammar and lexicon based on the Lexical-Functional Grammar (LFG) formalism (Bresnan, 2001). The other is the implementation and operation of a comprehensive open treebanking environment for building, hosting and exploring treebanks, thereby overcoming problems of maintenance and fragmentation due to treebanks being scattered and dependent on various platform-dependent software.

The project has recently cooperated with META-NORD (Vasiļjevs et al., 2012) (2011–2013), in the Information and Communication Technologies Policy Support Programme (CIP ICT-PSP), aimed at creating an open infrastructure to promote the accessibility and reuse of language resources and technologies (LRT). The META-NORD consortium included organizations from all the Nordic and Baltic countries. Among its recent results has been the documentation, rights clearance, licensing and sharing of many language resources, including treebanks, via the META-SHARE[1] catalogue and repository, thereby making LRT more readily available to R&D.

While the details of this cooperation are presented elsewhere,[2] the present paper will summarize the present state of the INESS infrastructure with a focus on functionality and usability.

2 Hosting treebanks in the INESS infrastructure

INESS currently provides the most comprehensive web-based treebanking services available. A normal web browser is sufficient as a client platform for accessing, searching and downloading treebanks, and also for the annotation of LFG-based parsebanks, including computer-aided manual disambiguation, text cleanup and handling of unknown words (Rosén et al., 2009, 2012b). The search functionality has recently been extended and simplified (Meurer, 2012). Search, visualization, resource management and cataloguing have been streamlined and work in similar ways for treebanks in several different paradigms (LFG, constituency and various dependency formats), thus simplifying access to a variety of resources.

For these reasons, INESS has become an attractive service for research groups who have developed or want to develop treebanks, but who cannot or do not want to invest in their own suite of web services for treebanking. Since the project start in 2010, INESS has been hosting an increasing number of treebanks, small and large. Among the larger treebanks developed by others and made available through INESS are the Icelandic Parsed Historical Corpus (IcePaHC, 73,014 sentences) (Wallenberg et al., 2011), the German Tiger treebank (50,472 sentences with dependency annotation, 9,221 with LFG annotation) (Brants et al., 2002) and the dependency part of the Bulgarian BulTreeBank (11,900 sentences) (Simov and Osenova, 2004). There is also a collection of sizable treebanks for Northern Sami (in total more than 1.7 million sentences, although not manually checked).

[1] `http:/meta-share.tilde.lv`

[2] A paper about the cooperation between INESS and META-NORD has been accepted for the Workshop on *Nordic Language Resources Infrastructure (NoLaReIn)* at NoDaLiDa 2013.

Figure 1: Choosing treebanks in INESS with the parallel treebanks setting.

The INESS infrastructure also offers tools for the development and exploitation of *parallel* corpora. Two parallel corpora aligned at sentence level were recently constructed and documented in cooperation with META-NORD. One of these is based on excerpts of a number of translations of the novel *Sofies Verden* (*Sophie's World*) (Gaarder, 1991), annotated for different languages and various formalisms, amounting to 26 aligned language pairs. Annotations of translations of a document from the Acquis communautaire were also aligned, currently yielding 21 pairs.

The ParGram project (Butt et al., 2002) has recently started using INESS as a testbed for semi-automatically aligning several LFG treebanks at phrase level. For treebanks constructed with parallel LFG grammars, the technology developed in the XPAR project (Dyvik et al., 2009) makes it possible to automatically align c-structure phrases based on manually indicated translational equivalences between f-structures. The resulting parallel treebanks are currently too small for exploitation, but their construction and exploration is an important proof of concept and a test of the parallel grammar construction endeavor in ParGram.

A screenshot with an overview of the treebank selection interface, with the option of choosing parallel treebanks, is shown in Figure 1. This figure also shows the 30 languages for which there are currently treebanks in the INESS infrastructure.

3 Accessing treebanks in the INESS interface

When accessing treebanks, users may want to identify treebanks and their provenance, for instance to correctly cite the materials. Access to many resources also requires that users be authenticated and authorized to use the materials under specified conditions. In recent joint work with META-NORD more attention has therefore been paid to the documentation and licensing of treebanks in INESS.

Relevant information is presented in the user interface after login. By way of example, Figure 2 shows how information on the Sofie Estonian treebank is presented the first time an authenticated user accesses this resource in INESS. A text describing the terms of the license and other metadata is presented to the user. The user can then choose to accept this license by clicking on the "Accept" button. This procedure is only necessary for treebanks with restrictions; in many cases the restrictions amount to no more than the requirement of attribution.

Resource: META-NORD Sofie Estonian Treebank

Description:

The Estonian part of the META-NORD Sofie Parallel Treebank.

This is a syntactically annotated parallel corpus based on the first chapters of the novel "Sofies verden" (Sophie's World) by Jostein Gaarder, published by Aschehoug forlag. The treebank consists of grammatical annotations of extracts from the Estonian translation of the novel, originally created as part of the Nordic Treebanking Network and now included in the extended META-NORD Sofie Parallel Treebank. The Estonian translation is published by Koolibri Publishing House. For more information, see the metadata description of the META-NORD Sofie Parallel Treebank.

ACCESS TO THE TREEBANK

The following terms hold for the use of the treebank:

The IPR holdership remains with Jostein Gaarder, who kindly permits INESS to distribute the "Sofie analyses" outside the project under the following terms of use:

a. The "Sofie analyses" can only be used for language technology research and development.

b. The users of the "Sofie analyses" are not allowed to redistribute or to publish the "Sofie analyses", only the knowledge and work that has been made on the basis of the "Sofie analyses",

c. The users of the "Sofie analyses" will ensure appropriate acknowledgement/references to the author of the original text, Jostein Gaarder, to Aschehoug Publishing House, Koolibri Publishing House and to the project INESS.

The alignments in the META-NORD Sofie Parallel Treebank are available under a CC-BY license (http://creativecommons.org/licenses/by/3.0/).

Attribution text:

"Alignments provided by the project INESS (www.iness.uib.no) in cooperation with META-NORD (http://www.meta-nord.eu/)."

Availability: available-restrictedUse

License: See «Description» for details
Restrictions of use: Treebank creators and IPR holders must be attributed, see «Description» for details.
Access medium:
Can be used for academic and commercial purposes.

By accepting the terms of the license you will be granted access to the resource.
Accept

Funding project (Nordic Treebank Network)
Project URL: ***http://w3.msi.vxu.se/~nivre/research/nt.html***
Project funded by: the Nordic Language Technology Program
Project start date: , end date:

Language: Estonian (***ISO code:*** et)

Figure 2: Presentation of documentation and user license, Sofie Estonian treebank.

4 The INESS Norwegian treebank

One of the main activities of the INESS project is the development of a large treebank for Norwegian, obtained by parsing automatically with an LFG grammar on the XLE parsing platform. The Norwegian treebank is growing and consists of a number of different genres in fiction and non-fiction. Part of the treebank is being efficiently manually disambiguated with the LFG PARSEBANKER (Rosén et al., 2009) Currently 4568 Norwegian sentences (46735 words) have been manually (at least partially) disambiguated; of these, 3602 sentences (35450) have been fully disambiguated. Based on the increasing number of manually disambiguated and quality controlled sentences, a stochastic disambiguator has been implemented which currently operates on the fly for any new sentences that are added.

(1) *Men det var helt umulig.*
But it was completely impossible.

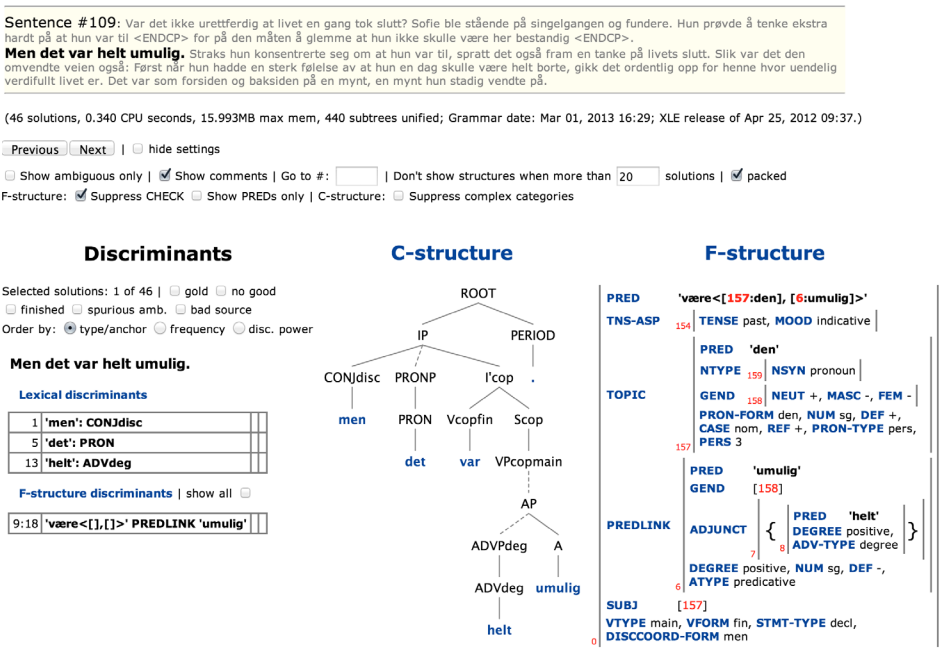

Figure 3: Screenshot of a disambiguated sentence in the INESS Norwegian treebank.

Figure 3 shows the manually disambiguated analysis of the sentence in example 1. Parsing resulted in 46 analyses for this sentence, and 211 discriminants were calculated (see Rosén et al. (2007) for the use of discriminants in LFG). Of these 211, only four were chosen in order to select the intended analysis.

5 Conclusion and outlook

INESS is becoming established as an infrastructure hosting a variety of existing treebanks as well as actively promoting the development of new treebanks. The infrastructure provides added value to existing treebanks through an increasing number of services, now including streamlined documentation and metadata, cataloguing information, access and licensing procedures, search, visualization and download. The INESS infrastructure is fully functional.

In the future, access will be further improved through user authentication by means of single sign-on via federated identity servers (eduGain) and through the use of persistent identifiers to identify resources (including also complex resources and possibly parts of resources). The resources and services in INESS will also be catalogued and linked in CLARINO (the Norwegian part of the CLARIN network) and in the *Language Technology Resource Collection for Norwegian – Språkbanken*, hosted at the National Library of Norway.

Acknowledgments

The reported work has been funded by the Research Council of Norway, the University of Bergen, and the European Commission under CIP-ICT-PSP grant agreement no. 270899.

References

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories*, pages 24–41.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell, Malden, MA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan*.

Dyvik, H., Meurer, P., Rosén, V., and De Smedt, K. (2009). Linguistically motivated parallel parsebanks. In Passarotti, M., Przepiórkowski, A., Raynaud, S., and Van Eynde, F., editors, *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, pages 71–82, Milan, Italy. EDUCatt.

Gaarder, J. (1991). *Sofies verden: roman om filosofiens historie*. Aschehoug, Oslo, Norway.

Meurer, P. (2012). INESS-Search: A search system for LFG (and other) treebanks. In Butt, M. and King, T. H., editors, *Proceedings of the LFG '12 Conference*, LFG Online Proceedings, pages 404–421, Stanford, CA. CSLI Publications.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012a). An open infrastructure for advanced treebanking. In Hajič, J., De Smedt, K., Tadić, M., and Branco, A., editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Rosén, V., Meurer, P., and De Smedt, K. (2007). Designing and implementing discriminants for LFG grammars. In King, T. H. and Butt, M., editors, *The Proceedings of the LFG '07 Conference*, pages 397–417. CSLI Publications, Stanford.

Rosén, V., Meurer, P., and De Smedt, K. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Van Eynde, F., Frank, A., van Noord, G., and De Smedt, K., editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT7)*, pages 127–133, Utrecht. LOT.

Rosén, V., Meurer, P., Losnegaard, G. S., Lyse, G. I., De Smedt, K., Thunes, M., and Dyvik, H. (2012b). An integrated web-based treebank annotation system. In Hendrickx, I., Kübler, S., and Simov, K., editors, *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT11)*, pages 157–167, Lisbon, Portugal. Edições Colibri.

Simov, K. and Osenova, P. (2004). BulTreeBank Stylebook. BulTreeBank Project Technical Report 5, Bulgarian Academy of Sciences.

Vasiļjevs, A., Forsberg, M., Gornostay, T., Haltrup Hansen, D., Jóhannsdóttir, K., Lyse, G., Lindén, K., Offersgaard, L., Olsen, S., Pedersen, B., Rögnvaldsson, E., Skadiņa, I., De Smedt, K., Oksanen, V., and Rozis, R. (2012). Creation of an open shared language resource repository in the Nordic and Baltic countries. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, pages 1076–1083, Istanbul, Turkey. European Language Resources Association (ELRA).

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., and Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC) version 0.9.