META-NORD

Baltic and Nordic Branch of the European Open Linguistic Infrastructure

Project no. 270899

Deliverable D3.4
Parallel Treebanks

Version No. 1.0

January 31, 2013

Document Information

Deliverable number:	3.4
Deliverable title:	Parallel Treebanks
Due date of deliverable:	31 January 2013
Actual submission date of deliverable:	31 January 2013
Main Author(s):	Koenraad De Smedt, Gunn Inger Lyse, Gyri Losnegaard, Anje Gjesdal (UIB)
Participants:	UIB (task leader), TILDE, UCPH, UT, UHEL, HI, LKI, UGOT
Internal reviewer:	UHEL
Workpackage:	3
Workpackage title:	Enhancing language resources
Workpackage leader:	UCPH
Dissemination Level:	PU
Version:	1.0
Keywords:	treebanks, annotated corpora, parallel corpora

EXECUTIVE SUMMARY

Treebanks provide detailed grammatical analyses of language data and can be used for developing applications in language understanding, translation etc. This deliverable reports on Task 3.4, which has been aimed at improving the accessibility of treebanks and harmonising and linking treebanks across languages. META-NORD has, in cooperation with INESS, collected, annotated, indexed, aligned and documented treebanks, cleared their rights and has made them available to authenticated users for download as well as for interactive search and processing. This effort has mainly been concentrated on the Nordic and Baltic languages represented in the consortium. The treebanks are documented in metadata records at META-SHARE while the data are available for browsing, search, visualization and download from the INESS treebanking infrastructure.

Table of Contents

Abbreviations

Table 1 Abbreviations

Abbreviation	Term/definition
Acquis	Acquis Communautaire, JRC Acquis Multilingual Parallel Corpus of EU/EEA law texts and treebanks based on this material
CG	Constraint Grammar
DA	Depositor's agreement
HI	University of Iceland
INESS	Infrastructure for the Exploration of Syntax and Semantics
IPR	Intellectual property rights
LFG	Lexical-Functional Grammar
NTN	Nordic Treebank Network
R&D	Research and Development
Sofie	Sofies Verden (Gaarder 1991) and treebanks based on this material
SVG	Scalable Vector Graphics
UCPH	University of Copenhagen
UGOT	University of Gothenburg
UHEL	University of Helsinki
UIB	University of Bergen
UT	University of Tartu

1. Introduction

Treebanks are databases of detailed grammatical analyses of language data. They are especially useful for training stochastic disambiguating parsers and for parser induction. Therefore they are important tools in R&D on a wide range of language understanding applications. Parallel treebanks are constructed through alignment of monolingual treebanks at the sentence or phrase levels. They are useful in translation studies and in R&D on high quality machine translation. The aim of Task 3.4 has been to improve the accessibility of treebanks and to harmonize and link treebanks across languages. This deliverable reports on the methods used and results achieved in this task.

The development of high quality treebanks is labor intensive, specialised research and therefore a relatively costly undertaking for small languages. In order to promote the accessibility of existing treebanks for the languages in META-NORD, the project has concentrated on the annotation, harmonisation, curation, documentation and licencing of treebanks, and has improved the availability of tools to build, maintain, link, explore, filter and download treebanks in an open infrastructure optimised for these tasks.

In order to reach those goals, META-NORD has cooperated extensively with the *Infrastructure for the Exploration of Syntax and Semantics* (INESS[1]), a large infrastructure project in Norway. The INESS project runs from 2010 until 2016 and is funded by the Research Council of Norway and the University of Bergen. The INESS project goals are (1) to develop a large high quality treebank with deep syntactic analysis for Norwegian and (2) to host treebanks in a common infrastructure with a suitable interface for visualisation, search and processing. It is the first large infrastructure especially for treebanking, fully accessible through a web interface without the need to install any client software except for a web browser (Rosén et al 2012). The INESS project has provided an important cooperation on Task 3.4 in the META-NORD project. The treebank metadata are catalogued at META-SHARE while the data are available from INESS.

2. Monolingual treebanks

2.1. Selection of materials and annotations

The Norwegian novel *Sofies verden* (Gaarder 1991) was chosen as a suitable basis for parallel treebanking because it is linguistically rich and professionally translated in many languages, and because some monolingual treebanks already existed for text selections from this material in some languages in the META-NORD area. Previous work was done by the Nordic Treebank Network (NTN[2]), funded by the Nordic Language Technology Program (2001-2005) but had not been maintained and was no longer accessible. It was decided to gather those treebanks, document them, supplement them with additional new treebanks for some languages where this effort was feasible, and make the resulting resources accessible in a coherent way.

NTN annotation files for materials from *Sofies verden* were obtained from Tekstlaboratoriet at the University of Oslo, and treebanks for Danish, Estonian, German, Icelandic and Swedish were selected from this material. An English treebank was obtained

[1] http://iness.uib.no

[2] http://w3.msi.vxu.se/~nivre/research/nt.html

from SMULTRON.[3] Although Finnish and Norwegian NTN annotations were also available, these were not suitable. Instead, new treebanks were created for these languages: the FinnTreeBank team of UHEL developed small hand-annotated samples from the novel for the FinnTreeBank, while the INESS project contributed new LFG annotations for the Norwegian material. A treebank for the Georgian translation annotated by Paul Meurer was also included in the collection for the sake of linguistic diversity. INESS is open to the future addition of other language versions.

The *Sofie* treebanks made available through META-NORD show considerable diversity with respect to both language families covered and linguistic formalisms that are represented. *Sofie Danish Treebank* is a dependency treebank, semi-automatically annotated according to the guidelines used to create the Danish Dependency Treebank and automatically converted to TIGER-XML by the DTAG program. *Sofie Estonian Treebank* is a constraint grammar (CG) treebank, automatically parsed with a CG parser assigning syntactic function labels and enhanced with manually added constituencies. *Sofie Icelandic Treebank* is a constituency treebank which was manually annotated by the late Gunnar Hrafn Hrafnbjargarson. *Sofie Swedish Treebank* is a dependency treebank, automatically created with the Maltparser tool. *Sofie German Treebank* was annotated with the Annotate tool, followed by an automatic deepening of the flat syntax trees. *Sofie Finnish Treebank* is a manually annotated dependency-CG treebank created by the UHEL FinnTreeBank team for FinnTreeBank and META-NORD. *Sofie Norwegian Treebank* and *Sofie Georgian Treebank* are automatically parsed with LFG grammars developed in the NorGram and INESS projects, producing c-structures and f-structures; the analyses are manually (interactively) disambiguated by the use of discriminants and the treebanks are downloadable in Negra/Tiger XML format.

Furthermore, small pilot treebanks were constructed for the *JRC Acquis Multilingual Parallel Corpus of EU/EEA law texts*, which provides materials from a different genre. This corpus contains texts in all EU languages and also some non-official European languages, including all META-NORD languages, although not every document is available in all those languages. Given the diversity of domains in the whole *Acquis* and the resulting difficulties for vocabulary coverage, we selected only one document from the *Acquis* for this parallel treebank, a document of appropriate length and which is available in all the relevant META-NORD languages.

INESS also provided annotations of the English version of the selected *Acquis* document, which was parsed with an English LFG grammar and manually disambiguated via the INESS treebanking interface, in a similar way as was done for Norwegian.

Besides the *Sofie* and *Acquis* treebanks which were the basis of parallel treebanks, some other freestanding monolingual treebanks based on different sources were also selected in cooperation with the INESS project. These include treebanks for Finnish, Icelandic and Norwegian in the linguistic area of META-NORD as well as for a number of other languages inside and outside the linguistic area of META-NET (including smaller languages in the META-NORD geographical area such as Faroese and Northern Sami).

2.2. Rights clearance

UIB and its META-NORD partners have cleared the rights of the original and selected translations of *Sofies verden*. A depositor's agreement (DA) was signed with Aschehoug,

[3] Slightly different versions of the German and Swedish treebanks are also available in SMULTRON (http://www.cl.uzh.ch/research/paralleltreebanks/smultron_en.html)

the publisher of the Norwegian original, who also wrote a recommendation letter to the publishers of the translations, which were subsequently contacted. Signed depositor's agreements have so far been obtained for the Swedish, Estonian, Danish, Icelandic, German and Georgian translations. It has not been possible to negotiate an agreement for the Finnish translation because its translator and IPR holder does not wish to make the text available for research. Thus, the Finnish *Sofie* treebank will until further notice not be made publicly available. The DAs used for the *Sofie* materials are based on the standard META-SHARE template and have restrictions on the redistribution of the texts while allowing the use of the texts for R&D purposes in language technology, which is our most important purpose. Treebanks with restrictions can only be accessed by registered users who are logged in.

The rights of the annotations were cleared separately. The Danish, Estonian, German, Icelandic and Swedish analyses were all created in the Nordic Treebank Network, but except for the Estonian and German treebanks which had creation data encoded in the metadata part of the XML, it proved challenging to identify the creator(s) and rights holder of each individual annotation. In consultation with NTN network coordinator Prof. Joakim Nivre and NTN project co-workers Mathias Buch-Kromann and Kadri Muischnek (the creators of the Swedish, Danish and Estonian annotations respectively), it was agreed that Nivre should sign a DA for all annotations created within the NTN project, on behalf of the annotation group. These treebanks are for the time being licenced with the user terms of the source text. The rights of the Finnish and Norwegian annotations were cleared in individual agreements, under the CC-BY licence.

The source document for the *Acquis* treebanks, Directive 2002/74/EC, did not need any rights clearance because it is part of the *Acquis Communautaire* (the total body of EU law applicable in the the EU Member States), which is in the public domain. Permission to use the Norwegian translation was obtained from the Ministry of Foreign Affairs in Norway. Building on our experience with the *Sofie* annotations, where multiple annotations were developed in the same project, it was decided to create a common DA for all annotations created in META-NORD by META-NORD participants.

A CC-BY licence was also selected for the *Acquis* annotations, ensuring proper acknowledgement of the European Commission's Office for Official Publications (OPOCE), the Norwegian and Icelandic Ministries of Foreign Affairs that have translated the directive into their respective (non-EU) languages, and of the JRC that publishes the *JRC Acquis Multilingual Corpus*, which was used for document selection. The user is informed that if reproducing the text literally, the terms of use of the source text applies.

2.3. Treebank documentation: metadata collection

For the *Sofie* treebanks, it was a major challenge that NTN project results and documentation was no longer maintained and partly inaccessible. Many of the URLs on the project webpage were obsolete, and information about the treebanks and their creators were partly or completely missing. A small amount of metadata was found in some of the annotation files, and some information was available on the NTN webpages. Most of the information necessary to create adequate metadata descriptions for the treebanks and to ensure rights clearance with the creators of the annotations, was obtained by contacting former NTN network participants Joakim Nivre and Matthias Buch-Kromann. Buch-Kromann (previously Trautner Kromann) also uploaded missing documentation to the Copenhagen Dependency Treebank's Google code repository, including two HTML pages documenting the tools developed in NTN and the common representation formats chosen

(Tiger XML for treebanks). This documentation will be maintained via the treebank section of the INESS webpages.

For the treebanks, as well as for the Finnish and Norwegian *Sofie* annotations, the following metadata was supplied from the META-NORD contributing partners:

- Annotation mode (automatic, semi-automatic, manual);
- Grammar/parser (type or name of tool);
- Grammar origin/creator (project, persons);
- Grammar type/formalism (constituency, dependency, LFG, etc.);
- Output format (Tiger XML, CoNLL etc.);
- Tagset (documentation url, taglist, etc.);
- Name(s) of the annotator(s).

Validation of the treebanks by META-NORD partners is documented in section 3.

2.4. Integration in INESS

A text selection from the first chapters of the Norwegian original of *Sofies verden* was syntactically annotated with NorGram, an LFG grammar. The text was automatically parsed and manually disambiguated, supported by discriminants in the LFG Parsebanker (Rosén et al 2009). Annotated translations were also created or collected for Danish, Estonian, Finnish, Icelandic and Swedish. Furthermore, treebanks for translations into English, Georgian and German were also added even if these languages do not fall into the META-NORD linguistic area, simply because they were available and because they provide useful linguistic reference points. These *Sofie* and other treebanks were brought together in a uniform treebanking environment in INESS. Ingested treebanks are catalogued and indexed in order to allow efficient search. They can be browsed, inspected, filtered and downloaded. The server middleware was written in Common Lisp on top of an open source web server in the same language. The use of the same high level programming language throughout the whole system has resulted in a highly flexible system in which all annotation and analysis services are seamlessly integrated. The system is easy to modify at all levels, which promotes a fast evolution in response to user needs. The standard download format is NEGRA.

Figure 1 illustrates the top level overview of the available treebanks. By clicking on the choices for languages, collections and types, one can make a selection from the available treebanks.

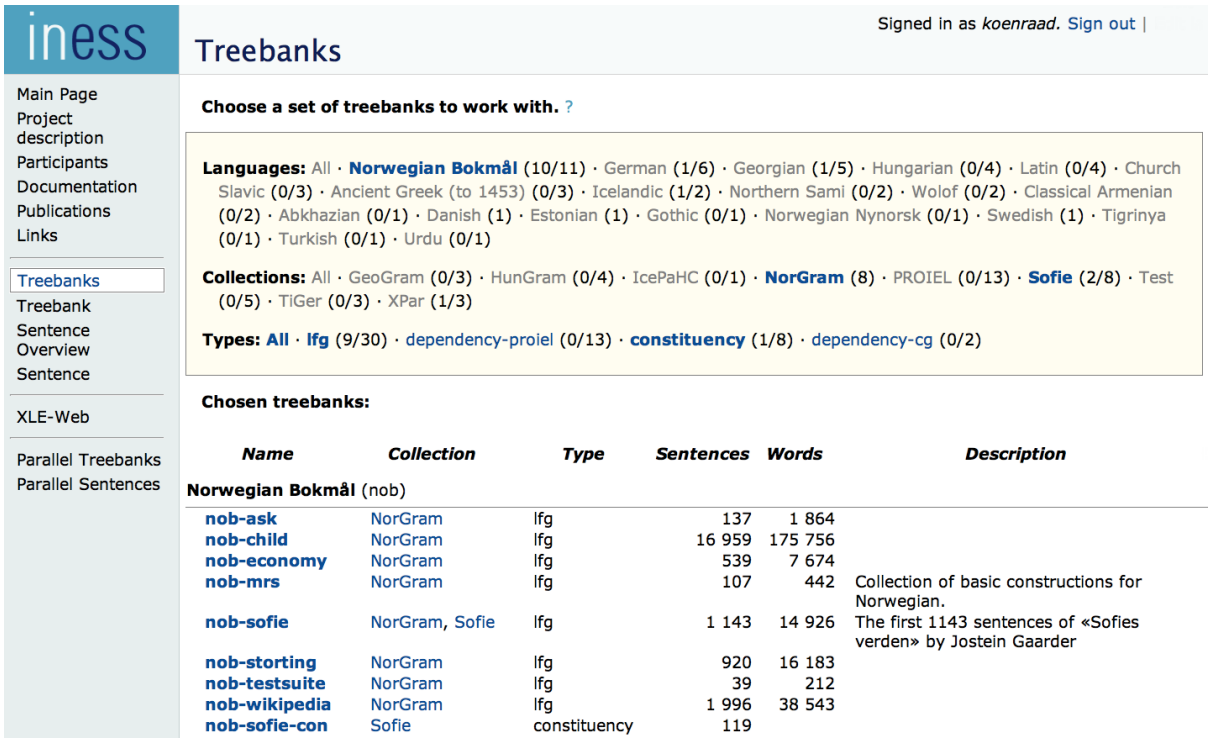

Figure 1 Overview screen in INESS with treebank selection tool

The visualisation of the Norwegian sentence *"Sofie Amundsen var på vei hjem fra skolen"* is shown in Figure 2. Visualisations are based on Scalable Vector Graphics (SVG), which is supported by modern web browsers. The presentation is adapted to the type of treebank. Currently the INESS middleware can handle LFG, constituency and dependency treebanks in various formats. The system also has powerful search facilities (Meurer 2012).

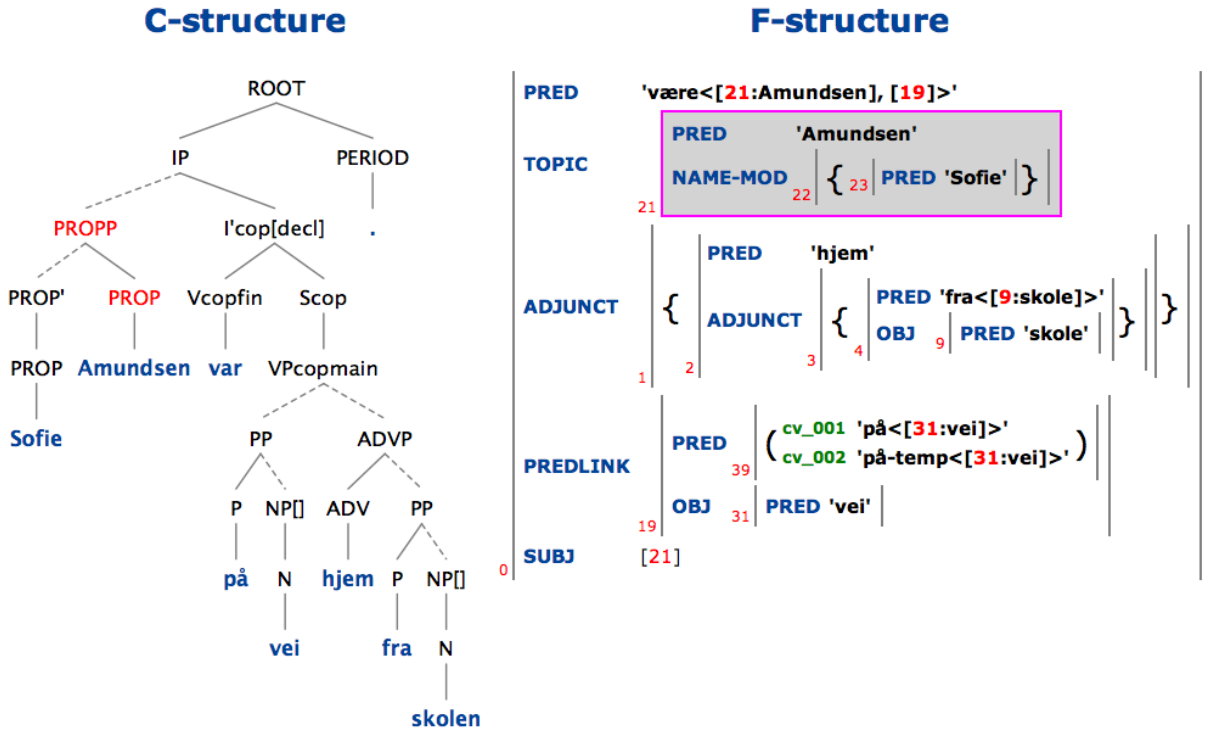

Figure 2 Visualisation of c-structure (tree, left) and f-structure (feature-value-graph, right)

The INESS project encourages owners of treebanks to use INESS as a virtual repository for their treebank(s). Grammar developers are also invited to upload their grammar or parser to INESS for use in online and interactive treebank development and alignment.

2.5. Lists of monolingual treebanks

The following tables list treebanks constructed by or obtained through the META-NORD partners. Their metadata are catalogued on META-SHARE while the treebanks themselves made available on INESS to registered users after login, due to licensing. Table A gives an overview of the monolingual treebanks collected and developed for the *Sofie* Parallel Treebank, and table B of the treebanks in the *Acquis* Parallel Treebank. Table C describes other treebanks obtained or identified by META-NORD in the action on horizontal treebanking. Table D lists other Norwegian treebanks developed in cooperation with INESS.

In these tables, size is measured in sentences. The status field indicates whether IPR and licence issues are resolved. If rights have been cleared for a resource, it is available for download and for searching and viewing through the INESS interface. Until rights are cleared for a given resource, it is only accessible for inspection through the web interface. For *Sofie*, all rights have been cleared except for the Finnish translation.

A. META-NORD Sofie Parallel Treebank

Treebank	Lang	MN-lang	Size	Type	Origin	Status
Sofie Danish Treebank	da	yes	102	dependency	NTN	available for download and through interface
Sofie Estonian Treebank	et	yes	52	CG	NTN	available for download and through interface
Sofie Finnish Treebank	fi	yes	506	dependency-CG	META-NORD/ FinnTreeBank	finished but not publicly available
Sofie Icelandic Treebank	is	yes	73	constituency	NTN	available for download and through interface
Sofie Norwegian Treebank	nb	yes	255	LFG	META-NORD/ INESS	available for download and through interface
Sofie Swedish Treebank	sv	yes	215	dependency	NTN	available for download and through interface
Sofie English Treebank	en	no	528	constituency	SMULTRON	available for download and through interface
Sofie Georgian Treebank	ka	no	1.025	LFG	INESS	available for download and through interface
Sofie German Treebank	de	no	225	constituency	NTN	available for download and through interface

B. META-NORD Acquis Parallel Treebank

Treebank	Lang	MN-lang	Size	Type	Origin	Status
Acquis Danish Treebank	da	yes	102	dependency	META-NORD project	available for download and through interface
Acquis Estonian Treebank	et	yes	78	dependency-CG	META-NORD project	available for download and through interface
Acquis Finnish Treebank	fi	yes	122	dependency	META-NORD project	available for download and through interface
Acquis Icelandic Treebank	is	yes	73	constituency with some dependency features	META-NORD project	available for download and through interface
Acquis Norwegian Treebank	nb	yes	100	LFG	META-NORD project	available for download and through interface
Acquis Swedish Treebank	sv	yes	102	dependency	META-NORD project	available for download and through interface
Acquis English Treebank	en	no	94	LFG	META-NORD	available for download and through interface

C. Other treebanks and treebank-related resources obtained through the META-NORD partners or META-NORD related dissemination

Treebank	Lang	MN delivery	Size	Type	Origin	Status
INESS Sofie Norwegian Treebank	nb	UIB, Batch 3	1.000 sentences	LFG	META-NORD/INESS	available for download and through interface
Icelandic Parsed Historical Corpus (IcePaHC)	is	HI, Batch 1	73.014 sentences	constituency	The Icelandic Treebank	available for download and through interface
The Dependency Part of BulTreeBank	bg	UIB, Batch 3	196.000 tokens	HPSG	BulTreeBank project	available for download and through interface
The Morphologically Annotated Part of BulTreeBank	bg	UIB, Batch 2	214.000 tokens	HPSG	BulTreeBank project	available for download and through interface
FinnTreeBank3	fi	UHEL, Batch 3	170.000 tokens	NA	FinnTreeBank project	downloadable from http://www.ling.helsinki.fi/ under a CC BY 3.0 licence, will also be made available in INESS
Turku Dependency Treebank	fi	UHEL, Batch 3	6.303 words, 88.418 tokens	dependency	The Turku BioNLP Group	downloadable from http://bionlp.utu.fi/fintreebank-finnish.html under a CC BY-SA 3.0 licence, will also be made available in INESS
Norwegian translations of Acquis Communautaire (aligned on document level)	no	UIB, Batch 3	NA	NA	META-NORD project	parallel corpus, available for download at http://www.nb.no/Tilbud/Forske/Sprakbanken/Tilgjengelege-ressursar/Tekstressursar

The Finnish treebanks listed under C are available under CC-BY licences and may also be made available via INESS. Other treebanks provided by META-NORD partners, such as the Copenhagen Dependency Treebank, the Estonian Treebank and the Swedish Treebank, may also be considered for distribution via INESS. However, since the resources are already available elsewhere, the integration of these resources in INESS has not been

given priority. Negotiations are still ongoing to include treebanks from partners in the other projects in the META-NET family, such as a treebank for the Romanian *Sofie* translation.

Only a small part of the Norwegian translations of the *Acquis Communautaire* is currently syntactically and morphologically annotated. However, UIB has aligned the whole collection of Norwegian translations with the materials of the *JRC Acquis Multilingual Corpus* at document level, identified and classified documents in the language varieties Bokmål and Nynorsk, and transferred the collection to the META-NORD collaborator Språkbanken in Norway, where it has been made available for download, while further enhancement of the material such as text cleanup, XML encoding and paragraph alignment with the multilingual collection is being considered. The material remains a candidate for further syntactic annotation in INESS, but the experience from the small *Acquis* Parallel Treebank suggests that the document structure (with many tables, lists and enumeration), and the complexity of legal terminology and syntax, makes it difficult to analyse them automatically.

D. Norwegian treebanks under construction and partly available in INESS

Treebank	Size (sentences)	Origin
The nno-child treebank	3.597	INESS/Children's books in Norwegian Nynorsk from the The National Library of Norway
The nno-novel treebank	7.513	INESS/Novels in Norwegian Nynorsk from the The National Library of Norway
The nob-ask treebank	137	INESS/Essays collected from language tests, taken from the ASK language learner corpus of Norwegian as a second language
The nob-child treebank	266.131	INESS/Children's books in Norwegian Bokmål from the National Library of Norway
The nob-economy treebank	539	NA
The nob-mrs treebank	107	INESS/LOGON
The nob-newspaper treebank	6.323	INESS/NNC
The nob-pargram treebank	37	INESS/PARGRAM
The nob-snl treebank	6.125	INESS/Store Norske Leksikon
The nob-storting treebank	920	NA
The nob-testsuite treebank	39	INESS/LOGON
The nob-wikipedia treebank	1.997	INESS/Norwegian Wikipedia

A considerable amount of material in the INESS Norwegian Treebanks has been analysed and is currently available. The selection and annotation of more text material for the INESS Norwegian Treebanks is in progress. Upon completion of the INESS project, these treebanks will encompass approximately 50 million automatically annotated and disambiguated sentence analyses. The Norwegian treebanks in INESS are automatically parsed with the LFG grammar NorGram, which is undergoing continuous development. 5000 sentence analyses are being manually disambiguated and will form the basis for a statistical disambiguation of the remaining analyses, and the manual labour contributes to an improvement of the grammar's coverage, since shortcomings identified during the manual disambiguation are reported by the annotators and lead to necessary updates of grammar and lexicon. The sizes reported in table D (and to a certain extent also tables

A-C) indicate the number of sentences, and not necessarily the number of parsed sentences, since some sentences may not have been parsed due to coverage problems, limits on parser processing capacity, etc. It is, however, advisable to keep these in the database, in particular for monolingual treebanks that have been aligned.

Several treebanks for Norwegian and for other languages outside the META-NORD area are also available but are not described in detail here. These include other EU languages, non-EU languages and old Indo-European languages, among others Abkhazian, Faroese, Georgian, Hungarian, Northern Sami, Polish, Tamil, Turkish, Urdu, and Wolof (see the INESS website for a full overview).

3. Parallel treebanks

3.1. Sofie and Acquis parallel treebanks

META-NORD has delivered two parallel treebanks, accessible through a uniform web interface and state-of-the-art search tools. These parallel treebanks were created by linking monolingual treebanks across languages at sentence level. Alignments were made for the META-NORD languages Danish, Estonian, Finnish, Icelandic, Norwegian and Swedish, as well as for English. However, alignments with Finnish *Sofie* are currently not publicly available due to the aforementioned lack of permission from the translator who is the IPR holder. The META-NORD *Sofie* Parallel Treebank has also been extended with German and Georgian. The attempt to create a pilot parallel treebank with phrase alignments of Norwegian and Danish LFG annotations was not successful, mainly due to the low coverage of the Danish LFG grammar.

The *Sofie* monolingual treebanks (see 2.5, table A) were linked across languages by alignment at sentence level. This alignment was performed pairwise, i.e. between pairs of sentences in two languages; this process was performed manually, supported by a computational alignment tool provided by INESS. The result is a series of bilingual alignments between treebanks in different languages (see list in section 3.4).

The annotated *Acquis* materials have been aligned in the same way as the *Sofie* treebanks (see list in section 3.4).

3.2. Parallel treebanks in INESS

Sentence alignments are represented in a stand-off fashion, in a separate XML file which lists explicit links between an identifier of a sentence of the source text and an identifier of the corresponding sentence of the target text. The following is an extract for the alignment of the Danish and Icelandic *Acquis* treebanks. This example shows that the first, second and third Danish sentences are all aligned with the first Icelandic sentence, and that the fourth Danish sentence is aligned with the second Icelandic sentence.

```
<?xml version="1.0" encoding="utf-8"?>
<aligned-treebanks source-treebank="dan-jrc-acquis-con" source-type="dependency-cg"
target-treebank="isl-jrc-acquis-con" target-type="constituency">
  <link source-id="s1" target-id="s1"/>
  <link source-id="s2" target-id="s1"/>
  <link source-id="s3" target-id="s1"/>
  <link source-id="s4" target-id="s2"/>
 ...
```

</aligned-treebanks>

The parallel treebanks are documented on META-SHARE, while the alignment files can be downloaded from the INESS site; the parallel treebanks can also be browsed, searched and visualised online using the INESS interface, as shown in Figure 3, where syntactic structures for corresponding sentences in two different languages are presented side by side.

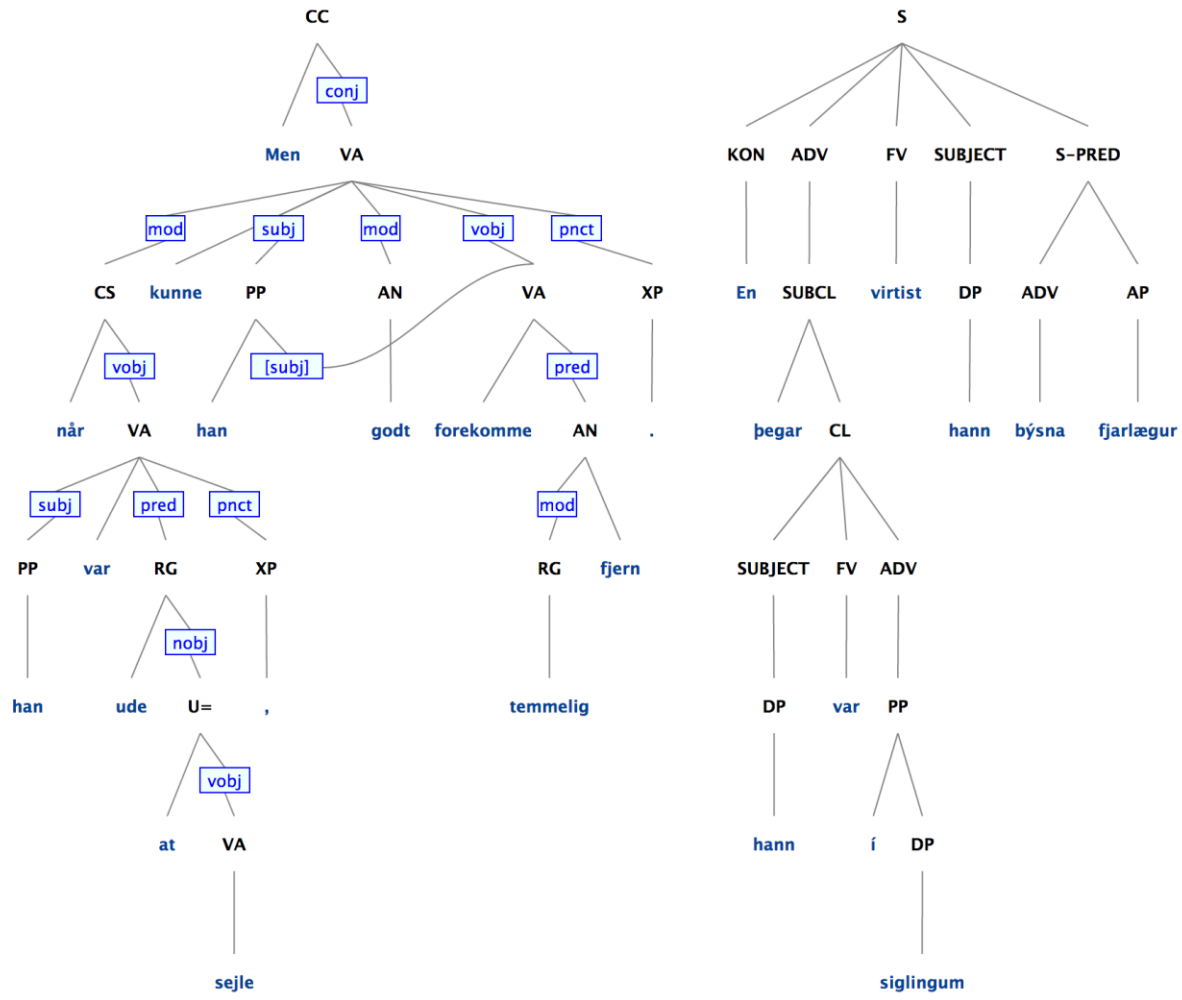

Figure 3 Illustration of visualisation of aligned Danish (left) and Icelandic (right) structures

3.3. Metadata representation of parallel treebanks

Parallel treebanks are complex in two different respects: they are composed of several monolingual treebanks, which are complex in their own right, with both a text component and one or more layers of annotation. Because the infrastructure still does not handle the description of complex resources, the representation of parallel treebanks in META-SHARE has been challenging. The implementation of a scheme for representing such resources was envisioned at the META-SHARE-forum in October 2011, but was not accomplished, and several approaches were considered to meet this problem.

One option was to register the parallel treebank as one resource and add a separate "corpusTextInfo" part for each language module. However, since the "lingualityInfo" and the "languageInfo" properties are described in the "textInfo" part, the multilinguality

dimension of the resource is then lost. The same problem holds if creating one metadata record for each monolingual treebank, since these are monolingual by nature and there is no good way of describing the fact that in combination, they represent a multilingual resource. The proposed solution from the META-SHARE developers was to create one metadata record, using the "sizePerLanguage" to give the minimal information for each language. Given the diversity of our monolingual treebanks and the number of properties we want to describe, in particular for the *Sofie* Parallel Treebank, which is an alignment of existing and new monolingual treebanks of different origin and types, this solution would not sufficiently meet our metadata requirements.

We finally opted for a resource description with one "mother" (multilingual, parallel) metadata record, and one for each of its (monolingual) components. The metadata records are linked using the "relation" feature. The monolingual treebanks are related to the "mother" resource with a "part of" relation, and to its sister resources with an "alignedWith" relation. The Parallel treebank is related to its "children" with a "hasPart" relation, and is further described as a multilingual text corpus which also lists its language components in the "sizeInfo" part.

For the future, we would like a metadata schema for parallel treebanks to support:

- one metadata record with individual descriptions of the monolingual treebank components;
- for each monolingual component (i.e., treebank), individual descriptions of its layers (i.e., source text and annotations);
- validation of each monolingual treebank, and the possibility to document the number of acceptable analyses, not acceptable analyses, unparsed sentences etc. (as well as which sentences or parse units this information holds for).

Other requests for improvement of META-SHARE that were reported, but not met, are the representation of language varieties in Norwegian and the provision of templates (minimal schemas) for different types of resources.

3.4. List of parallel treebanks

The following bilingually aligned parallel treebanks based on the *Sofie* and *Acquis* materials are currently available for download and browsing, with the exception of the Finnish *Sofie* treebank which has been retracted (cf. section 2.2).

1	Sofie Danish Treebank ⇄ Sofie German Treebank
2	Sofie Danish Treebank ⇄ Sofie Georgian Treebank
3	Sofie Danish Treebank ⇄ Sofie Finnish Treebank *(not publicly available)*
4	Sofie Danish Treebank ⇄ Sofie Estonian Treebank
5	Sofie Danish Treebank ⇄ Sofie Icelandic Treebank
6	Sofie Danish Treebank ⇄ Sofie English Treebank
7	Sofie Danish Treebank ⇄ Sofie Norwegian Treebank
8	Sofie Danish Treebank ⇄ Sofie Swedish Treebank
9	Sofie German Treebank ⇄ Sofie Norwegian Treebank

10 Sofie German Treebank ⇄ Sofie Georgian Treebank
11 Sofie German Treebank ⇄ Sofie Swedish Treebank
12 Sofie German Treebank ⇄ Sofie Icelandic Treebank
13 Sofie German Treebank ⇄ Sofie English Treebank
14 Sofie German Treebank ⇄ Sofie Finnish Treebank *(not publicly available)*
15 Sofie German Treebank ⇄ Sofie Estonian Treebank
16 Sofie English Treebank ⇄ Sofie Icelandic Treebank
17 Sofie English Treebank ⇄ Sofie Georgian Treebank
18 Sofie English Treebank ⇄ Sofie Norwegian Treebank
19 Sofie English Treebank ⇄ Sofie Swedish Treebank
20 Sofie English Treebank ⇄ Sofie Estonian Treebank
21 Sofie English Treebank ⇄ Sofie Finnish Treebank *(not publicly available)*
22 Sofie Estonian Treebank ⇄ Sofie Icelandic Treebank
23 Sofie Estonian Treebank ⇄ Sofie Georgian Treebank
24 Sofie Estonian Treebank ⇄ Sofie Swedish Treebank
25 Sofie Estonian Treebank ⇄ Sofie Finnish Treebank *(not publicly available)*
26 Sofie Finnish Treebank ⇄ Sofie Icelandic Treebank *(not publicly available)*
27 Sofie Finnish Treebank ⇄ Sofie Georgian Treebank *(not publicly available)*
28 Sofie Finnish Treebank ⇄ Sofie Norwegian Treebank *(not publicly available)*
29 Sofie Finnish Treebank ⇄ Sofie Swedish Treebank *(not publicly available)*
30 Sofie Icelandic Treebank ⇄ Sofie Georgian Treebank
31 Sofie Icelandic Treebank ⇄ Sofie Swedish Treebank
32 Sofie Icelandic Treebank ⇄ Sofie Norwegian Treebank
33 Sofie Georgian Treebank ⇄ Sofie Swedish Treebank
34 Sofie Norwegian Treebank ⇄ Sofie Swedish Treebank

35 Acquis Danish Treebank ⇄ Acquis Finnish Treebank
36 Acquis Danish Treebank ⇄ Acquis Swedish Treebank
37 Acquis Danish Treebank ⇄ Acquis English Treebank
38 Acquis Danish Treebank ⇄ Acquis Icelandic Treebank
39 Acquis Danish Treebank ⇄ Acquis Norwegian Treebank
40 Acquis Danish Treebank ⇄ Acquis Estonian Treebank
41 Acquis English Treebank ⇄ Acquis Swedish Treebank
42 Acquis English Treebank ⇄ Acquis Finnish Treebank
43 Acquis English Treebank ⇄ Acquis Estonian Treebank
44 Acquis English Treebank ⇄ Acquis Icelandic Treebank
45 Acquis English Treebank ⇄ Acquis Norwegian Treebank

46	Acquis Estonian Treebank ⇄ Acquis Norwegian Treebank
47	Acquis Estonian Treebank ⇄ Acquis Finnish Treebank
48	Acquis Estonian Treebank ⇄ Acquis Swedish Treebank
49	Acquis Estonian Treebank ⇄ Acquis Icelandic Treebank
50	Acquis Finnish Treebank ⇄ Acquis Norwegian Treebank
51	Acquis Finnish Treebank ⇄ Acquis Icelandic Treebank
52	Acquis Finnish Treebank ⇄ Acquis Swedish Treebank
53	Acquis Icelandic Treebank ⇄ Acquis Norwegian Treebank
54	Acquis Icelandic Treebank ⇄ Acquis Swedish Treebank
55	Acquis Norwegian Treebank ⇄ Acquis Swedish Treebank

3.5. Validation

All contributing parties have been asked, to the extent possible, to check the alignments and evaluate the syntactic annotations for their languages. The results are described for each META-NORD partner. The *Acquis* English Treebank was evaluated by UIB. The English, Georgian and German *Sofie* annotations were not evaluated. Links to relevant documentation have been provided in the META-SHARE metadata records of the English and German Sofie treebanks respectively.

UCPH

Validator: Jürgen Wedekind

The Sofie Danish Treebank (103 sentences/parse units): All annotations and alignments validated.

The Acquis Danish Treebank (102 sentences/parse units): Approximately 50% of the analyses are judged acceptable, but it has not been recorded which sentences this holds for. All alignments have been manually verified.

UT

Validator: Kadri Muischnek

The Sofie Estonian Treebank (52 sentences/parse units): All annotations and alignments validated.

The Acquis Estonian Treebank (78 sentences/parse units): The dependency CG parser output has been manually corrected, and all alignments have been manually verified.

UIB

Validator: Gyri Smørdal Losnegaard

The Sofie Norwegian Treebank (255 sentences/parse units): Good analysis: 225 (88%), no good: 3 (1%), no analysis: 19 (7%), acceptable analysis 8 (3%).

The Acquis Norwegian Treebank (102 sentences/parse units): Good analysis: 50%, no good: 12%, no analysis: 24%, acceptable analysis 13%

The Acquis English Treebank (94 sentences/parse units):

Good analysis: 39 (42%), no good: 23 (24%), no analysis: 9 (10%), acceptable analysis 23 (24%).

All alignments have been manually verified.

UHEL

Validator: Kristiina Muhonen

The Sofie Finnish Treebank (506 sentences/parse units): 13 (3%) acceptable OCR errors, 9 (2%) unacceptable analyses.

The Acquis Finnish Treebank (122 sentences/parse units): 17 (14%) have unacceptable or no analyses.

HI

Validator: Kristín M. Jóhannsdóttir

The Sofie Icelandic Treebank (194 sentences/parse units): Several sentences are missing.

The Acquis Icelandic Treebank (73 sentences/parse units): The linguistic marking is satisfactory, but for 13 sentences annotation features have entered into the terminal nodes (the text). This can probably be cleaned up manually.

UGOT

Validator (alignments): Malin Ahlberg

The Sofie Swedish Treebank (215 sentences/parse units) and the Acquis Swedish Treebank (102 sentences/parse units): All alignments have been manually checked. The annotations have not been evaluated, but validation documentation has been added to the metadata record in the form of an article evaluating the parsing model, reporting a labeled attachment score (both the head and the label are correct) of approx. 77/100 (Nivre 2006).

4. Conclusion and sustainability

A successful demonstration of harmonisation and linking of treebanks across languages and improving their accessibility has been achieved in Task 3.4. This action ought to be maintained and expanded in order to be fully effective. The INESS project continues to provide a specialised repository and laboratory for treebanking, which is run and upgraded until at least 2016 and will then be maintained and preserved in the context of the CLARINO project (the Norwegian part of the CLARIN infrastructure). The INESS infrastructure is open to languages also outside of META-NORD.

All treebanks made available through META-NORD have been documented on META-SHARE, while the resources themselves are distributed through the dedicated treebanking platform INESS, because the META-SHARE platform does not have adequate facilities for treebanking purposes, while INESS provides a rich treebanking environment allowing the user to view visualizations of different types of treebanks, view parallel sentences and analyses, search for syntactic structures in and across treebanks etc., in addition to downloading annotation and alignment files. It is a priority for UIB to make users aware of the advantages of specialized infrastructures, so that the user may, for instance, judge the quality of a treebank by viewing it before download, or to download only relevant sections as needed.

The construction and exploitation of treebanks are complicated processes. The handling of monolingual treebanks is currently well supported for a range of grammar formalisms. Multilingual treebanks have been successfully aligned at sentence level, while alignment at phrase level would be a next major step. This has been a topic of research (Adesam 2012,

Dyvik et al. 2009) but this cannot be fully realised until a major parallel grammar construction effort is undertaken for several languages.

The representation of metadata for complex resources such as parallel treebanks has been problematic in META-SHARE. Tilde has offered to share their prototype for linking external resource portals (such as EuroTermBank) with META-SHARE. This linking makes it possible to dynamically list each specific resource in META-SHARE, and can be applied to other language resource specific portals such as INESS.

The clearance of rights for the materials was time consuming and in one case could not be resolved satisfactorily, so that the Finnish treebank is not publicly available. This case illustrates our dependence on the cooperation of authors and publishers and the need to initiate clearance procedures early, before spending efforts on annotation.

A supporting action aimed at awareness raising and dissemination related to the Horizontal Action on Treebanking was the organisation of a META-RESEARCH Workshop on Advanced Treebanking at LREC on May 22, 2012 in Istanbul. This workshop was organised together with participants from the METANET4U, CESAR and T4ME projects. The workshop was successful and proceedings were published (Hajič et al 2012; cf. also D5.2.2).

5. References

Adesam, Yvonne. The Multilingual Forest: Investigating High-quality Parallel Corpus Development. PhD dissertation, Department of Linguistics, Stockholm University, 2012.

Helge Dyvik, Paul Meurer, Victoria Rosén, and Koenraad De Smedt. Linguistically motivated parallel parsebanks. In Marco Passarotti, Adam Przepiórkowski, Sabine Raynaud, and Frank Van Eynde, editors, Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories, pages 71–82, Milan, Italy, 2009. EDUCatt.

Jostein Gaarder. Sofies verden: Roman om filosofiens historie. Aschehoug, 2012.

Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco (eds.) META-RESEARCH Workshop on Advanced Treebanking at LREC2012, Istanbul, Turkey, May 22, 2012.

Paul Meurer. INESS-Search: A search system for LFG (and other) treebanks. In Miriam Butt and Tracy Holloway King, editors, Proceedings of the LFG ’12 Conference, pages 404–421, Stanford, CA: CSLI Publications, 2012.

Joakim Nivre. Inductive Dependency Parsing. Springer, 2006.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. An open infrastructure for advanced treebanking. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco (eds.) META-RESEARCH Workshop on Advanced Treebanking at LREC2012, pages 22–29, Istanbul, Turkey, May 2012.

Victoria Rosén, Paul Meurer, and Koenraad De Smedt. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank Van Eynde, Anette Frank, Gertjan van Noord, and Koenraad De Smedt, editors, Proceedings of the Seventh Inter- national Workshop on Treebanks and Linguistic Theories (TLT7), pages 127–133, Utrecht. LOT, 2009.

Gyri Smørdal Losnegaard, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad

De Smedt, Helge Dyvik, and Paul Meurer. What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. In Jan Hajič, Koenraad De Smedt, Marko Tadić, and António Branco, editors, META-RESEARCH Workshop on Advanced Treebanking at LREC2012, pages 69–76, Istanbul, Turkey, May 2012.