# Baltic and Nordic Parts of the European Linguistic Infrastructure

*Inguna Skadiņa[1], Andrejs Vasiļjevs[1], Lars Borin[2], Krister Lindén[3], Gyri Losnegaard[4], Sussi Olsen[5], Bolette S. Pedersen[5], Roberts Rozis[1], Koenraad De Smedt[4]*

(1) Tilde, Vienibas gatve 75a, Riga, Latvia
(2) University or Gothenburg, Box 100, 405 30 Gothenburg, Sweden
(3) University of Helsinki, Unioninkatu 40, Helsinki, Finland
(4) University of Bergen, Postboks 7800, 5020 Bergen, Norway
(5) University of Copenhagen, CST, Njalsgade 140, 2300 Copenhagen S

metanord@tilde.lv

ABSTRACT

This paper describes scientific, technical, and legal work done on the creation of the linguistic infrastructure for the Nordic and Baltic countries. The paper describes the research on assessment of language technology support for the languages of the Baltic and Nordic countries, work on establishing a language resource sharing infrastructure, and collection and description of linguistic resources. We present improvements necessary to ensure usability and interoperability of language resources, discuss issues related to intellectual property rights for complex resources, and describe extension of infrastructure through integration of language-resource specific repositories. Work on treebanks, wordnets, terminology resources, and finite-state technology is described in more detail. Finally, our approach on ensuring the sustainability of infrastructure is discussed.

KEYWORDS: language resources and tools, linguistic infrastructure, under-resourced languages, multilinguality, treebanks, wordnets, terminology banks.

# 1 Introduction

The previous decades of research in language technologies (LT) have produced numerous language resources and tools for languages of the Nordic and Baltic countries. At the same time, not only are there major gaps in availability of the key resources, but the existing ones are often hard to find and difficult to use. Resources are dispersed among different institutions and local repositories, and they are often coded in proprietary formats lacking interoperability and uniformity. There are also restricted or unclear intellectual property rights. These factors are major stumbling blocks for the development and research of language technology.

To overcome these difficulties, the Nordic and Baltic countries play a leading role in pan-European activities regarding the creation of the European open linguistic infrastructure. Major progress is achieved by the CLARIN [1] (Váradi et al., 2008) initiative creating a language resource infrastructure for research in humanities. Another complementary infrastructure is under development by the META-NET network[2] focusing on the practical needs of developers, users, and researchers of multilingual resources.

The Baltic and Nordic countries are active participants in both — CLARIN and META-NET — networks. Official languages of these countries (Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian, and Swedish), as well as other languages spoken in these countries, are under-resourced in respect to availability of at least some of the key language technologies or resources. Thus, such initiatives as CLARIN and META-NET are essential for identification, documentation, and making widely available the language resources for these languages, and these initiatives provide significant support in filling essential gaps in resources and technologies.

In the META-NET network, work on the creation of an open European linguistic infrastructure and making language resources widely available was performed through four projects supported by the European Commission: CESAR (Central and Eastern European countries), METANET4U (Southern European countries and the United Kingdom), META-NORD (Baltic and Nordic countries), and T4ME (initiator and coordinator of META-NET network). In this paper, we describe our experience and work done on building the Baltic and Nordic parts of the European linguistic infrastructure in the META-NORD project (Vasiļjevs et al., 2011; Skadiņa et al., 2011), which included the following major tasks:

- Research on the language technology landscape in the Nordic and Baltic countries in terms of language use, language technology and resources, and main actors.
- Work on identification and collection of language resources in the Baltic and Nordic countries and documenting, processing, linking, and upgrading them to agreed standards and guidelines.
- Research and practical work on consistent approaches, practices, and standards that ensure wider accessibility, easier access, and reuse of quality language resources.
- Research and development work done on linking and validating the Nordic and Baltic wordnets, harmonisation of multilingual Nordic and Baltic treebanks, consolidation of multilingual terminology resources across European countries, and development of mature and language independent tools.
- Issues related to intellectual property rights (IPR) for the sharing of language resources.

[1] Common Language Resources and Technology Infrastructure, http://www.clarin.eu
[2] http://www.meta-net.eu

- Work on building and operating broad, non-commercial, community-driven, inter-connected repositories, exchanges, and facilities that can be used by different categories of user communities.

## 2 Language technology landscape

Important steps towards the creation of a linguistic infrastructure are an assessment of the state of the art of the language technology field, identification of existing language resources, assessment of resources, and understanding the needs of potential users. Reports on the national language technology landscape[3] for each official language spoken in the Nordic and Baltic geographical area (Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, both Norwegian varieties Nynorsk and Bokmål, and Swedish) describe and analyse the situation of the language community and the position of the language service and language technology industry. These reports also contain general facts about the language (number of speakers, official status, dialects, etc.), particularities of the language, recent developments in the language, language cultivation, language in education, international aspects, as well as the role of the language on the Internet.

The reports also present assessment of language technology support and the core application areas of language and speech technology (e.g., language checking, web search, speech interaction, machine translation, etc.), and describe the situation with respect to these application areas. For each language, language resources and tools (LRT) are assessed based on the following criteria: quantity, availability, quality, coverage, maturity, sustainability, and adaptability. The results indicate that only with respect to the most basic tools and resources, such as tokenisers, PoS taggers, morphological analysers/generators, reference corpora, and lexicons/terminologies, the status is reasonably positive for all languages. However, tools for information extraction, machine translation, and speech recognition, as well as resources such as parallel corpora, speech corpora, and grammar, are rather simple and have limited functionality for some of the languages. For the most advanced tools and resources, such as discourse processing, dialogue management, semantics and discourse corpora, and ontological resources, most of the languages either have nothing of the kind, or their tools and resources have a quite limited scope.

This assessment of language technology support clearly demonstrates that for languages of the Baltic and Nordic countries, where the community of language resource creators and users is small, even a modest increase in the availability and quality of language resources is appreciable for technology developers, researchers, and end users.

In addition, detailed analysis and comparison of language technologies and resources across 30 European languages was carried out within the framework of META-NET. The comparison presents the situation for four key areas: machine translation, speech processing, text analysis, and resources. This study puts three small languages of the Nordic and Baltic region – Icelandic, Latvian, and Lithuanian – in the last cluster, defined as having major gaps for all of the four key areas (see Table 1). The relative ranking of the remaining five languages is slightly higher, although none of them comes close to the "larger" languages (English, French, Spanish, and German). "Moderate" support is provided only for Finnish in speech technologies and for Swedish with respect to language resources.

Besides objective limitations in the size of the LRT creator and user community, there are other obstacles which put some languages in the "upper" cluster, while others remain in the "bottom".

[3] These reports, also called Language White Papers, are available at http://www.meta-net.eu/whitepapers.

Among them is a lack of continuity in research and development funding. For instance, due to limited funding, Latvian language technology support does not reach the quality and coverage not only of that for English, but also for many under-resourced languages of the Baltic and Nordic region with a smaller number of speakers. In many cases, targeted national research and development activities are urgently needed to fill LRT gaps.

| **Speech processing** | | | | |
|---|---|---|---|---|
| **Excellent** | **Good** | **Moderate** | **Fragmentary** | **Weak/None** |
| | English | Czech, Dutch, **Finnish,** French, German, Italian, Portuguese, Spanish | Basque, Bulgarian, Catalan, **Danish, Estonian,** Galician, Greek, Hungarian, Irish, **Norwegian,** Polish, Serbian, Slovak, Slovene, **Swedish** | Croatian, **Icelandic, Latvian, Lithuanian,** Maltese, Romanian |
| **Machine Translation** | | | | |
| **Excellent** | **Good** | **Moderate** | **Fragmentary** | **Weak/None** |
| | English | French, Spanish | Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian | Basque, Bulgarian, Croatian., Czech, **Danish, Estonian, Finnish,** Galician, **Icelandic,** Irish, **Latvian, Lithuanian,** Maltese, **Norwegian,** Portuguese, Serbian, Slovak, Slovene, **Swedish** |
| **Text Analysis** | | | | |
| **Excellent** | **Good** | **Moderate** | **Fragmentary** | **Weak/None** |
| | English | Dutch, French, German, Italian, Spanish | Basque, Bulgarian, Catalan, Czech, **Danish, Finnish,** Galician, Greek, Hungarian, **Norwegian,** Polish, Portuguese, Romanian, Slovak, Slovene, **Swedish** | Croatian, **Estonian, Icelandic,** Irish, **Latvian, Lithuanian,** Maltese, Serbian |
| **Resources** | | | | |
| **Excellent** | **Good** | **Moderate** | **Fragmentary** | **Weak/None** |
| | English | Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, **Swedish** | Basque, Bulgarian, Catalan, Croatian, **Danish, Estonian, Finnish,** Galician, Greek, **Norwegian,** Portuguese, Romanian, Serbian, Slovak, Slovene | **Icelandic,** Irish, **Latvian, Lithuanian,** Maltese |

TABLE 1: Availability of LRT for languages of the Baltic and Nordic countries[4].

The need for large amounts of data and the complexity of language technology systems make it vital to develop both an open infrastructure and a more coherent research cooperation in order to spur greater sharing and reuse of language resources.

[4] The table is also available at http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

## 3 META-SHARE infrastructure in the Baltic and Nordic countries

For distribution and sharing of language resources, the distributed online platform META-SHARE (Piperidis, 2012) is used. It consists of independent META-SHARE nodes set up in different countries and interlinked into a federated repository. This freely accessible distributed online infrastructure provides facilities for describing, storing, preserving of language resources, and making them publicly available. Among various language data that can be considered useful for different purposes, META-SHARE places a strong focus on language data that are important in language technology development for building applications that are useful to EU citizens, primarily in their everyday communication and information search needs. META-SHARE is intended for providers and users of language resources and technologies such as LT developers, researchers, students, translators, technical writers and others.

Currently, META-SHARE nodes are set at the following organisations in the Baltic and Nordic countries: Tilde[5] (Latvia), University of Gothenburg[6] (Sweden), University of Helsinki[7] (Finland), Institute of Lithuanian Language[8] (Lithuania), University of Copenhagen[9] (Denmark), Norwegian National Library[10] (Norway), and University of Tartu[11] (Estonia). According to the architecture of META-SHARE, these nodes are networked, and the content of the individual LR repository is harvested into the managing META-SHARE node, which for the META-NORD consortium is set at Tilde. In a managing node, information about catalogued language resources is collected and synchronised with other managing nodes across Europe, thus providing access to the full catalogue of the pan-European infrastructure[12].

Besides META-SHARE repositories, we have a natural interest to integrate into our infrastructure several existing collections and databases of specific linguistic resources, such as term banks and treebanks. These repositories are collections of language resources, where each individual resource is a candidate to be listed in the META-SHARE catalogue. This could be done manually by entering all resource descriptions in the META-SHARE editor or by exporting the metadata from the respective repository, converting it into META-SHARE compliant schema (Gavrilidou et al., 2012), and importing into META-SHARE node. However, such approaches are time-consuming and need regular manual updates.

Our proposed and implemented solution for this infrastructure is to integrate complex linguistic resources or repositories of resources by adapting them to relevant data access and sharing specifications and interlinking them with META-SHARE. This means that a language resource-specific repository could seamlessly become a part of the META-SHARE network by enabling the harvesting of metadata through the META-SHARE communication protocol and ensuring the mapping of the respective data categories.

For this approach to work the metadata model of the language resource-specific node must include all the data categories that are mandatory in the META-SHARE repository, as well as include additional attributes required for the synchronization, such as unique ID, creation and modification date and revision number.

---

[5] http://metashare.tilde.com/
[6] http://spraakbanken.gu.se/metashare/
[7] http://metashare.csc.fi/
[8] http://meta-share.lki.lt/
[9] http://metashare.cst.dk/
[10] http://metashare.nb.no/
[11] http://metashare.ut.ee/
[12] META-SHARE described in details in Piperidis (2012).

The integration of a language resource-specific node with META-SHARE is implemented via proxy: it connects to a META-SHARE managing node just like any other META-SHARE node – the LR metadata provider is proxied to the rest of the META-SHARE network. Integration of a language resource-specific node with META-SHARE allows users to access a specific resource located on a remote repository directly via a link supplied to META-SHARE.

META-NORD project piloted extension of the META-SHARE infrastructure with resource-specific nodes by integrating distributed terminology database EuroTermBank as described in the Section 5.3.

## 4 Identification, collection, and description of language resources in META-SHARE

During the last two years, more than 500 resources and tools have been identified and made available by the META-NORD consortium. These include a broad range of different resources for different languages and language pairs that are suitable for a range of different LT purposes. Statistical breakdown of the language resources for META-NORD languages available from META-SHARE platform is summarized in Table 2.

| Danish | Estonian | Finnish | Icelandic | Latvian | Lithuanian | Norwegian | Swedish |
|---|---|---|---|---|---|---|---|
| 68 | 67 | 86 | 70 | 69 | 38 | 47 | 151 |

TABLE 2: Language resources and tools for META-NORD languages documented in META-SHARE platform.

The criteria for the selection of resources included: availability, popularity, suitability of resources for technology and product/application development, fitness for multilingual purposes, longevity, quality, and extensibility.

Initially, the main focus has been on written resources, however, recently there has been an increased effort to also include a certain number of audio/video resources and tools. Table 3 shows the distribution of tools and resources available through META-SHARE for the Baltic and Nordic languages.

| | |
|---|---|
| Lexical resources (excl. wordnets & speech) | 211 |
| Corpora (excl. treebanks & speech) | 182 |
| Tools (excl. speech) | 43 |
| Language description (grammars) | 2 |
| Treebanks | 31 |
| Resources for speech | 28 |
| WordNets | 12 |
| **Total** | **509** |

Table 3: META-NORD tools and resources identified and made available through META-SHARE.

Considerable work documenting, processing, linking, and upgrading these resources to agreed standards and guidelines has been performed as described in sections 4.1 and 4.2 below.

Activities dealing with multilingual wordnets, treebanks, and terminology (cf. Sections 5.1–5.3) provide examples of how these goals of interoperability in content have actually been achieved for several of the provided resources.

## 4.1 Metadata model for the description of language resources

Bird and Simons (2003) have proposed a very useful taxonomy for discussing language documentation and description. Although originally framed in the context of linguistic description, their 'seven dimensions of portability' are quite useful also when discussing language resource interoperability for language technology purposes. Here, language resource metadata and widely accessible metadata repositories are absolutely essential, addressing – directly or indirectly – four of their seven dimensions, namely *discovery*, *access*, *citation*, and *rights*.

All of the contributed language resources are consistently identified and described using a metadata standard developed as a part of the META-SHARE initiative (Gavrilidou et al. 2011, 2012; Desipri et al. 2012). The standard, which draws heavily on experiences from previous efforts – such as the OLAC (Bird and Simons 2001) and CLARIN CMDI (Broeder et al. 2010) metadata schemas – defines a minimal set of descriptors which must be specified for any resource, but also allows for supplying much richer information if desired, including free-form narrative descriptions. It is published in the form of an XML schema. Using the META-SHARE metadata editor makes it easy to describe resources using the META-SHARE metadata schema, since the editor, for example, provides listings of the controlled vocabularies of the metadata schema.

Uploading a large number of diverse language resources and tools and the concomitant creation or conversion of the associated metadata by several large European projects have constituted a kind of empirical acid test of the appropriateness and usefulness of the META-SHARE schema. We are happy to report that it has passed this test successfully. For the 'baseline' cases, metadata creation has on the whole been straightforward and unproblematic, and the META-SHARE community online support[13] has generally been able to help in many of the more non-obvious cases.

Of course, it has inevitably turned out that the schema does not cover everything. For instance, it was discovered that it caters less well for complex multilingual resources, such as the mixed-language variety text corpora collected in Norway (with articles written in the two official written Norwegian varieties, Bokmål and Nynorsk, in the same newspaper), as well as for the multilingual treebanks and linked wordnets (see section 5). This issue, in fact, touches on several design decisions made for the META-SHARE metadata schema at the very beginning, having to do, for example, with how language information is encoded and with the structuring of the metadata records themselves. Because of the work in CESAR, METANET4U, META-NORD, and T4ME projects, the issues arising in connection with the description of complex language resources have been brought to the forefront and will be addressed in future releases of the META-SHARE metadata schema.

## 4.2 Improving usability and interoperability of language resources

An important focus of our work has been on enhancing the interoperability of language resources and tools by upgrading selected resources to agreed standards. The upgrading activities have included the following:

[13] http://www.meta-share.org/portal/ (Note that for some reason all issues listed there are labelled "unanswered", although most of them actually have answers which can be seen if you click on the heading of an issue to open the discussion thread for that issue.)

**Improvement of resource documentation**, both formally structured (as META-SHARE metadata) and narrative documentation. For many resources, the narrative documentation has been much improved; the partners have in several cases spent a considerable amount of time writing and improving resource documentation, as well as – in practice, very important for wider interoperability – on producing documentation in English. Also, for increased user-friendliness, an XSL stylesheet (developed in the CESAR sister project) is routinely used to automatically convert META-SHARE metadata into a more human-readable form - from XML into a more human-readable textual rendering (which however does not add any information to that already present in the XML metadata).

**Technical format conversion**, e.g., a proprietary corpus format into TEI (Text Encoding Initiative[14]) or an idiosyncratic lexicon format into LMF[15] (Lexical Markup Framework; ISO 24613:2008; Francopoulo et al. 2006). Several partners have converted their lexical resources into LMF, e.g., the STO Danish dictionary lexical database, the SweFN++ Swedish lexical macro resource, the Norwegian SCARRIE lexicon, and the Lithuanian Standard language lexical database. Terminology resources, such as the Icelandic Term Bank and UHR's Termbase for Norwegian higher education institutions, have been converted into TBX[16] format (Term Base eXchange; ISO 30042:2008; Melby 2012).

Most of the corpus resources uploaded are now available in TEI-compatible formats. A specific example of how this format harmonisation has enhanced interoperability is the relative ease[17] with which the open-source Korp corpus processing and presentation platform, developed in Sweden at the University of Gothenburg (Borin et al. 2012)[18], has been deployed in Finland by the University of Helsinki for their Finnish corpora[19].

**Content model conversion/mapping/linking**, e.g., harmonising POS tagsets among corpora, or linking word senses among lexical resources with different sense granularities. The Danish STO lexicon, the Swedish lexicons developed at Språkbanken, University of Gothenburg, and Swedish corpus annotations have been partly linked to the ISOCAT DCR (Data Category Registry; ISO 12620:2009; Windhouwer and Wright 2012), although no explicit attempt has been made to use the same categories across the languages, except in the specific cases discussed below in sections 5.1–5.3. Several partners (e.g., the Icelandic and Swedish partners) have taken advantage of their work on harmonising content and upgraded their corpus resources, adding consistent linguistic annotations to corpora that previously were either unannotated or used several different annotation schemes.

Ensuring that language resources and tools adhere to standard formats is a necessary prerequisite for resource interoperability. As always, the proof of the pudding is in the eating, and the outcomes of investing a limited, quite reasonable amount of time on horizontal action lines (see section 5 below) – where selected resources have been interlinked between languages of the Nordic and Baltic countries – demonstrate clearly that the upgrading activities have successfully achieved their objectives.

[14] http://www.tei-c.org
[15] http://www.lexicalmarkupframework.org/
[16] http://www.ttt.org/oscarStandards/tbx/
[17] The characterisation "relative ease" is impressionistic, based on a comparison with practical experience from earlier attempts to deploy another piece of corpus software both at Gothenburg and Helsinki, for which obviously much less effort had been spent on issues of interoperability and modularity during its development.
[18] http://spraakbanken.gu.se/korp/#lang=eng
[19] http://korp.csc.fi/#lang=en

As a means to upgrade existing resources in the META-NORD consortium to agreed standards, various lexical resources have been upgraded to the Lexical Markup Framework (LMF). One example is University of Gothenburg's 21 lexical resources which have all been upgraded to LMF. As reported in Borin et al. 2012, this lexical infrastructure has one primary lexical resource SALDO as a pivot to which all other resources are linked.

Upgrading some of the more semantically orientated resources to LMF turned out to be quite complicated. The main obstacle being that, even though the framework contains mechanisms for specifying semantic information, the model is based upon the assumption that lexical entries are formal entities expressing one or more senses, not semantic entities having one or more formal realisations. For example, the Swedish Framenet has the semantic frame as the natural conceptual unit, but to be able to fit the information into LMF, the frame had to be split into several entries. Searching these resources can be done from the web page http://spraakbanken.gu.se/karp, and all of the resources can be downloaded from http://spraakbanken.gu.se/eng/resources/lexicon.

Another example is the multilingual dictionary ISLEX[20] with modern Icelandic as the source language and Danish, Norwegian, and Swedish as target languages. ISLEX has been upgraded to LMF in the META-NORD project by the Icelandic partner, University of Iceland (cf. Helgadóttir & Rögnvaldsson forthcoming). When converting multilingual dictionaries into LMF format, a special record would usually be made for each sense of every word in all the languages of the dictionary. A so-called "Sense Axis" would then be used to link closely related senses in different languages. For ISLEX, however, another course was taken: special records were made for each sense of the words in the source language, Icelandic, which in turn had translations for that sense in each of the target languages.

Finally, it should be mentioned that the upgrade of the Danish lexical database STO[21] (built on PAROLE[22]) has been completed by the University of Copenhagen. This database provides an 'intensional' morphological description meaning that each word form is not explicitly listed anywhere but the lexical entry is associated with a morphological pattern. In other words, the word forms are created on demand. Even if an intensional morphological description is possible in LMF, an extensional description of the morphology, where all word forms are created when dumping data from database to LMF, was chosen as the most user-friendly approach. In some aspects, the predefined LMF schemata did not correspond with the structure of the linguistic information in the STO, and various kinds of generalisation or nested information had to be expressed in new ways by means of features which have made it easier to get an overview of the lexical entries. Furthermore, the LMF nomenclature is much closer to general linguistic terminology. In LMF for example, STO data categories such as 'description' and 'construction' are called 'subcategorisation frame' and 'syntactic argument' respectively, concepts much easier for a user to comprehend.

## 4.3 Approach to intellectual property rights

Collecting Intellectual Property Rights (IPR) for a corpus often requires that permissions are acquired from multiple parties in order to make copies of the copyrighted parts of the corpus. Copies for personal use are often readily available from many sources but the right to distribute

20 The ISLEX dictionary can both be searched and downloaded from http://www.malfong.is/index.php?pg=islex&lang=en.
21 A Danish lexical database developed by the University of Copenhagen, Centre for language technology (Braasch et al. 2004). Samples and documentation of the STO-LMF lexicon can be found at http://cst.ku.dk/english/sto_ordbase/.
22 PAROLE was an EU project running from 1996-1998 with the aim of creating large, generic, and re-usable Written Language Resources for all EU Languages.

such copies is not (see, e.g., Oksanen and Lindén (2012)). In this section, we look at a practical solution for multilingual treebanks in META-SHARE created within the project.

The right to distribute linked multilingual treebanks requires a right to copy and distribute a collective work, so the most practical solution is that one party collects the rights to copy and distribute all parts of the work, after which that party can sign for all the others. This is not that much different from collecting a corpus of several books and their annotations. In our case, the multilingual corpora were open-source or the rights already allowed distribution via META-SHARE, so the only work that needed to be given an explicit right to be copied and distributed were the hand-made annotations and cross-lingual links. Since the cross-lingual linking of the multilingual treebanks was done during the project, the ownership of the links belonged to the project parties. Only the annotations that had been done before the project or with separate funding needed to be provided with an explicit license for META-SHARE to distribute them. For the treebanks, we chose CC-BY as the common license.

## 5 Horizontal action lines

Besides identification and description of linguistic resources in the Baltic and Nordic countries, a particular focus in our work was on the following linguistic resources and tools: treebanks, wordnets, terminology, and finite-state techniques. We call this work 'horizontal action lines', as these activities focus on a particular resource type and aim to harmonise, link across languages, and make these resources available through a common interface.

### 5.1 Treebanks

The horizontal action on treebanking has been aimed at improving the accessibility of treebanks and harmonising and linking treebanks across languages. Special regard was taken to the Nordic and Baltic languages which are under-resourced in this respect, but the action was also a truly open initiative inclusive of other languages and addressing needs of other META-NET members.

Efforts were focused on the annotation, harmonisation, curation, documentation, and licensing of treebanks. To reach these goals, we collaborated extensively with the INESS project,[23] which provided state-of-the-art tools based on an advanced server-based solution and a user-friendly web interface for browsing, search, visualisation, and download. The resources which were made available by the action include a number of monolingual treebanks and two parallel treebanks based on multiple alignments of monolingual treebanks.

One notable outcome was the Sofie Parallel Treebank, based on the Norwegian novel Sofies verden (Gaarder 1991), which is linguistically rich and professionally translated in many languages. Some monolingual treebanks previously existed for text selections from this material, but not all were accessible. Annotations developed by the Nordic Treebank Network (NTN) were obtained from Tekstlaboratoriet (University of Oslo), and those that were deemed suitable were integrated in INESS. An additional new treebank was constructed for Norwegian. Annotations for the bigger European languages — German and English (the latter from SMULTRON24) were included for completeness, as well as a Georgian annotation by Paul Meurer. Intellectual property rights were cleared for both source texts and annotations, but for Finnish no permission has been obtained. Although the amount of annotated material varies between languages and is somewhat limited (52

[23] http://iness.uib.no
[24] http://www.cl.uzh.ch/research/paralleltreebanks/smultron_en.html

to 1052 sentences), the number of pairwise alignments is high. The Sofie Parallel Treebank currently consists of 26 aligned language pairs and is open to new languages.

A small parallel treebank was also constructed from a single document selected from the JRC Acquis Multilingual Parallel Corpus of EU/EEA law texts,25 which provides materials from a different genre in both the official EU languages and some non-official European languages. The Acquis Parallel Treebank is also rather small (73 to 122 sentences depending on the language), but it currently consists of 21 aligned language pairs. Even if the Sofie and Acquis treebanks are based on relatively small texts, they demonstrate the potential of aligning treebanks across many languages. These parallel treebanks are documented in META-SHARE and are available for download on the INESS website.

Besides the monolingual treebanks which form the basis of the two parallel treebanks, the INESS infrastructure provides access to other treebanks in several languages. Among treebanks in the linguistic area of the Baltic and Nordic countries, we mention the Icelandic Parsed Historical Corpus (73,014 sentences), FinnTreebank3 (170,000 tokens), Turku Dependency Treebank (88,418 tokens), the Faroese Parsed Historical Corpus (3,713 sentences), and the INESS Norwegian treebank (about 5000 analysed sentences and still expanding). Treebanks for other languages outside this linguistic area include Abkhazian, Bulgarian, Georgian, Hungarian, Northern Sami, Polish, Tamil, Turkish, Urdu, Wolof, and the classical languages (Ancient Greek, Church Slavic, Classical Armenian, Latin, and Gothic).

The different types of treebanks (Lexical-Functional Grammar, Dependency Grammar, Constituency Annotation) are accommodated in standard formats (TigerXML, CoNLL-X, CG3-dependency, Penn Treebank II bracketing, and XLE prolog). Treebanks can be explored (searched, browsed, and viewed), aligned with other treebanks, and downloaded, all from a uniform web interface. A screenshot is presented in Figure 1.

Drawing on the obtained experience, INESS will sustain the treebanking action and continue to pursue good practices for documentation and IPR clearance. We encourage treebank developers to clear rights with rights holders of source texts prior to annotation and promote the use of explicit licenses wherever possible.

The metadata description of treebanks remains problematic, in particular for parallel treebanks which are complex resources, since META-SHARE does not provide for adequate description of such resources. A more appropriate treatment of complex resources will need attention in future work.

The mechanisms for linking external resource portals with META-SHARE (described in Section 3) make it possible to dynamically list each specific resource in META-SHARE and can be applied to other language resource-specific portals such as INESS.

25 http://ipsc.jrc.ec.europa.eu/index.php?id=198

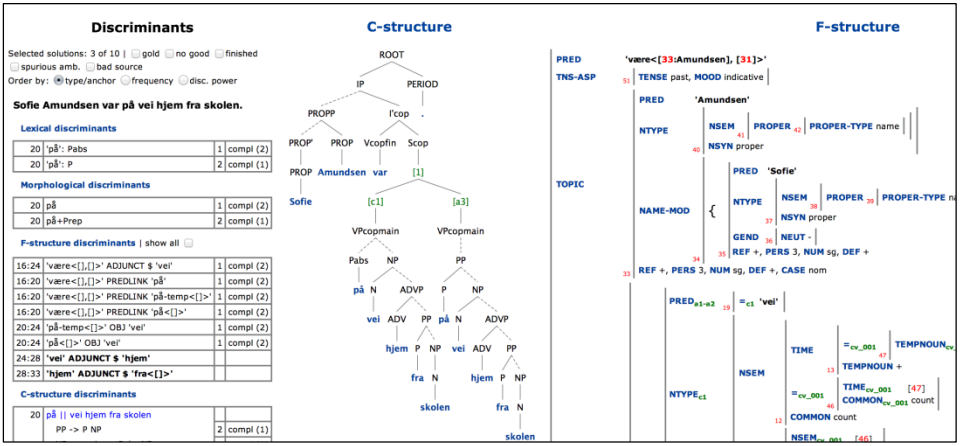


Figure 1: INESS web interface for treebanking, shown with the first sentence of the Norwegian Sofie treebank annotated in LFG.

## 5.2 Wordnets

The multilingual wordnet initiative has dealt with the pilot linking and validation of wordnets between the Nordic and Baltic languages. One central aim has been to perform a tentative comparison and validation of the linked wordnets, a related aim being to make the mono- and multilingual wordnets visible for further validation and comparison via a common web interface.

To this end, four pilot bilingual wordnets, each of 1,000 synsets, have been compiled using Princeton Core WordNet[26] as an Interlingua: Danish-Swedish, Danish-Finnish, Estonian-Finnish, and Finnish-Swedish.

In general, the semi-automatically linked wordnets are judged to be of a rather good quality, even if translations are not always 100% precise. An average of 2.2% errors and 7.0% slight mismatches have been reported from the individual validations. There does not appear to be a systematic bias to the errors. Though of course, some errors are systematically biased due to false friends, however, others seem to be just random errors. Most of the slight mismatches derive from diverging opinions on the understanding of *what a synset should contain*. Not surprisingly, wordnets that have been compiled via translations from Princeton WordNet have many senses per synset (just as Princeton WordNet), whereas wordnets that are monolingually compiled and based rather on synonymy registrations in conventional dictionaries have much less. A majority of the reported mismatches in the links are derived from exactly this discrepancy, since translations seem to be become more imprecise when a synset contains many word senses (Pedersen et al., 2013).

Apart from compiling the bilingual wordnets, all of the involved wordnets have undergone extensions and upgrades, including the Icelandic WordNet and the Norwegian WordNet, of which the latter has been developed from scratch during the project period (based on the Danish wordnet

[26] http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt

DanNet). All six monolingual wordnets and the four bilingual ones have been made available via META-SHARE[27], and the linked wordnets are viewable from the *WordTies*[28] web interface.

Future work includes an extension of the web interface to include more wordnets, as well as the generation of a broader comparison and validation of the wordnets included, i.e, broader than the ones provided via the 1,000 common links. We plan to include multilingual links for the full Core WordNet (5,000 synsets) for browsing and validation.

A broader comparison and validation of the wordnets would furthermore be fruitful and should be made feasible when the web interface is extended to include multilingual links for the full Princeton Core WordNet. Several discrepancies have been registered during the linking – other than the aforementioned different approach to the interpretation of the *synset* – such as discrepancies in taxonomical structure. Some have used an expert perspective, for example on animals, and thereby compiled a relatively deep taxonomy (i.e., the Finnish WordNet), whereas others have used a layman perspective adapted from a dictionary which is much more flat (i.e., the Danish wordnet). The number and selection of relations in the wordnets also differ; some have included only Princeton relations, others include EuroWordNet (Vossen, 1998) relations (i.e., Estonian and Danish), and others have adapted relations from other lexical projects, including qualia relations such as *used_for* and *made_by* relations (Danish and Norwegian WordNet).

## 5.3 Terminology resources

This task addressed a growing demand to consolidate distributed terminology resources across languages and domains and thus, has extended an open linguistic infrastructure with multilingual terminology resources.

Major progress in consolidating distributed terminology resources was accomplished by the EuroTermBank platform (Vasiļjevs and Schmitz, 2006). Our intention was not to duplicate resources stored on EuroTermBank, but to interconnect this content-specific Language Resource repository with META-SHARE. Using the approach described in the Section 3, we enabled harvesting of EuroTermBank metadata and integrated EuroTermBank within the META-SHARE infrastructure. This interlinking yielded 99 additional terminology resources now listed in META-SHARE. These resources themselves are available in EuroTermBank for online search.

As EuroTermBank provides specific facilities for searching, representing, and using terminology data, we use this platform for depositing new resources identified and collected in our work. As a result, EuroTermBank was extended from 2.3 to 2.8 million terms internally by adding 43 terminology collections to EuroTermBank – LKI (Lithuanian) terminology, EASTIN terminology of Assistive Technology, and 41 collection from Icelandic Termbank. Three other terminology collections remain in the negotiation stage. One external term base (BFT) was connected to EuroTermBank during this activity, two other (STRUNA — Croatian term bank, PIRARC – Multilingual database of Road Terms) remain in the negotiation stage. BFT is the Bank of Finnish Terminology in Arts and Sciences provided to META-SHARE by the University of Helsinki. This resource has been integrated for one-stop terminology search with EuroTermBank and, through interconnection of EuroTermBank with META-SHARE, is also listed in the META-SHARE catalogue.

The most outstanding result was achieved by Icelandic partners who succeeded at negotiating with 41 author whose work is contained in the Icelandic terminology bank. The rights to share for

[27] Note, however, that Norwegian WordNet actually includes two wordnets, one for Bokmål and one for Nynorsk.
[28] http://wordties.cst.dk

download were negotiated through a long and exhaustive process of negotiations. The resulting content was converted to the industry standard TBX format of terminology interchange, and both were uploaded to the META-SHARE network, as well as imported for centralised online lookup by users of the EuroTermBank terminology portal.

## 5.4 Finite-state techniques

To promote the development of language-independent natural language processing software, HFST–Helsinki Finite-State Technology[29] is included in CLARIN and META-NET. HFST is a framework for compiling and applying linguistic descriptions with finite-state methods. HFST currently collects some of the most important finite-state tools for creating morphologies and spellcheckers into one open-source platform and supports extending and improving the descriptions with weights to accommodate the modelling of statistical information. HFST offers a path from language descriptions to efficient language applications. Here, we have focused on making the HFST library available to the developers of new tools, new features in existing tools, or new language applications.

HFST is primarily designed for creating and compiling morphologies, which have been documented, for example, in Lindén et al. (2009; 2011; 2012). HFST contains open-source replicas of *xfst, lexc,* and *twolc*, which are well-known and well-researched tools for morphology building (see Beesley and Karttunen, 2003). The tools support both parallel and cascaded application of transducers.

There are a number of tools for describing morphologies. Many of them start with the item-and-arrangement approach in which an arrangement of sublexicons contains lists of items that may continue in other sublexicons. A formula for compiling such lexical descriptions was documented in Lindén et al. (2009). To realise the morphological processes, rules may be applied to the finite-state lexicon. In addition, HFST also offers the capability to train and apply part-of-speech taggers on top of the morphologies using parallel weighted finite-state transducers on text corpora (Lindén et al., 2012).

Using compiled morphologies, a number of applications have been created, e.g., spellcheckers for nearly 100 languages and hyphenators for approximately 40 different languages. The spellcheckers were derived from open-source dictionaries and integrated with OpenOffice and LibreOffice, e.g., a full-fledged Greenlandic spellchecker, which is a polyagglutinative language, is currently available for OpenOffice via HFST. By adding the tagger capability, we have also created an improved spelling suggestion mechanism for words in context (Lindén et al., 2012).

Some additional applications, such as synonym and translation dictionaries as well as a framework for recognising multi-word expressions for information extraction using HFST, have also been developed.

# 6 Conclusions and continuing activities

As described in this paper, a fully functional LR infrastructure is established in the Nordic and Baltic countries as a part of the pan-European META-NET network. This infrastructure will greatly support researchers, developers, and users providing information and access to variety of monolingual and multilingual resources for the languages of the Nordic and Baltic countries.

[29] http://hfst.sf.net

The cornerstone of the long-time viability of the developed infrastructure is the involvement of the following main national and/or regional actors:

- Producers, i.e., 'competence centres', typically public research centres, language institutes and academies, as well as private content owners like media companies and publishing houses.
- Aggregators, i.e., 'data centres' (repositories), usually supported by national and/or regional authorities.
- Sponsors, i.e., public authorities, research agencies, language councils, companies with language resource needs.
- Individual and institutional users from the research and industry sectors.

The META-SHARE infrastructure relies on the operation of interlinked META-SHARE nodes that are distributed and autonomously maintained by the participating institutions. As META-SHARE is a distributed platform, its sustainability depends on the willingness and ability of participating institutions to run META-SHARE nodes and to provide related services.

By signing letters of intent (indicating commitment, but not legally binding), we have committed to sustain the infrastructure by hosting and making available the META-SHARE repository of LRs through the META-SHARE network and providing technical and/or user support services for a period of at least 2 years.

In addition, the clearinghouse concept (a service centre for collection, classification and distribution of language resources) is considered for cooperation on long-term storage of resources with other similar service centres in other European countries, using cost-sharing principles. Inclusion of new actors and countries in the business model will be further elaborated in cooperation with the best practices of other META-SHARE centres. In most of the Baltic and Nordic countries, national centres promoting LR availability have already been established.

Specific plans for the curation of META-SHARE nodes and specific types of LRs are provided:

- Treebank resources are maintained and disseminated through the INESS project coordinated by the University of Bergen.
- Terminology resources are taken care of by EuroTermBank, which is maintained and supported by Tilde.
- The interlinked wordnet resources are maintained and disseminated nationally under coordination of the University of Copenhagen.
- In Norway, the National Library has set up a META-SHARE node and will continue to run it as a part of its efforts in CLARINO (the Norwegian CLARIN project).
- The University of Iceland is working on the formation of a national consortium, consisting of the University of Iceland, the Árni Magnússon Institute, the Reykjavik University, the National and University Library, and a few others, to maintain an Icelandic national repository of language resources.

## Acknowledgements

The META-NORD project has received funding from the European Commission through the ICT PSP Programme, grant agreement no. 270899. Many thanks to colleagues in META-NORD partner organizations: Jolanta Zabarskaitė from the Lithuanian language institute, Eiríkur Rögnvaldsson and Sigrún Helgadóttir from the University of Iceland and Kadri Vider from the University of Tartu.

## References


Beesley, K. and Karttunen, K. (2003). Finite State Morphology, CSLI publications.

Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education*, pages 7–18.

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language* 79(3), pages 557–582.

Borin, L., Forsberg, M., Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478.

Borin, L., Forsberg, M., Olsson. L., Uppström, J. (2012). The open lexical infrastructure of Språkbanken. *Proceedings of LREC 2012*, pages 3598-3602.

Braasch, A. and Olsen, S. (2004). STO: A Danish Lexicon Resource - Ready for Applications. In *Fourth International Conference on Language Resources and Evaluation*, Proceedings, Vol. IV. Lisbon, pages 1079-1082.

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C. (2010). A data category registry- and component-based metadata framework. In *Proceedings of LREC 2010*, pages 43–47.

Desipri, E., Gavrilidou, M., Labropoulou, P., Piperidis, S., Frontini, F., Monachini, M., Arranz, V., Mapelli, V., Francopoulo, G., Declerck, T. (2012). Documentation and user manual of the META-SHARE metadata model. http://www.meta-net.eu/meta-share/META-SHARE%20%20documentationUserManual.pdf.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). Lexical Markup Framework (LMF). *Proceedings of LREC 2006*, pages 233–236.

Gavrilidou, M., Labropoulou, P., Piperidis, S., Speranza, M., Monachini, M., Arranz, V., Francopoulo, G. (2011). Specification of metadata-based descriptions for language resources and technologies. T4ME deliverable D7.2.1. http://www.meta-net.eu/public_documents/t4me/META-NET-D7.2.1-Final.pdf .

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., Mapelli, V. (2012). The META-SHARE metadata schema for the description of language resources. *Proceedings of LREC 2012*, pages 1090–1097.

Helgadóttir, S., Rögnvaldsson, E. (forthcoming). Language Resources for Icelandic, *Workshop on Nordic Language Research Infrastructure*, NODALIDA 2013, Oslo.

Lindén, K., Silfverberg, M., Pirinen, T. (2009). HFST tool for morphology: An efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, Mahlow, C. and Piotrowski, M. (eds.). Berlin, Heidelberg: Springer Berlin Heidelberg, pages 28-47.

Lindén, K., Silfverberg, M., Axelson, E., Hardwick, S., Pirinen, T. (2011). HFST—Framework for Compiling and Applying Morphologies. In *Systems and Frameworks for Computational Morphology*. Mahlow, C. & Piotrowski, M. (eds.). Springer, Vol. 100, pages 67-85.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Silfverberg, M., Pirinen, T. A. (2012). Using HFST for Creating Computational Linguistic Applications. In *Computational Linguistics Applications*, Piasecki, M., and Przepiórkowski, A., Springer-Verlag.

Melby, A.K. (2012). Terminology in the age of multilingual corpora. *Journal of Specialized Translation* 18, pages 7–29.

Oksanen, V. and Lindén, K. (2012). Building shared language research environments inside the European Union: How to optimize the system based on experiences from real life. In *First Thematic Conference on the Knowledge Commons*. Louvain-la-Neuve, Belgium.

Pedersen, B. S., Borin, L., Forsberg, M., Kahusk, N., Lindén, K., Niemi, J., Nisbeth, N., Nygaard, L., Orav, H., Rögnvaldsson, E., Seaton, M., Vider, K., Kaarlo, V. (2013) Nordic and Baltic wordnets aligned and compared through "WordTies". In *Proceedings of Nodalida 2013* (in press).

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pages 36-42.

Skadiņa, I., Vasiljevs, A., Borin, L., de Smedt, K., Linden, K., Rognvaldsson, E. (2011). META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries. In *Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm (LRTS)*, Chiang Mai, Thailand, pages 107-114.

Váradi, T., Krauwer, S., Wittenburg P., Wynne, M., Koskenniemi, K. (2008). CLARIN: common language resources and technology infrastructure. In *Proceedings of the Sixth International Language Resources and Evaluation Conference.*

Vasiļjevs, A., Pedersen, B.S., de Smedt, K., Borin, L., Skadiņa, I. (2011). META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure. In *NODALIDA 2011 workshop Visibility and Availability of LT Resources, NEALT Proceedings Series,* Vol.13, pages 18-22.

Vasiļjevs, A. and Schmitz, K.D. (2006). Collection, harmonization and dissemination of dispersed multilingual terminology resources in an online terminology databank. In *Proceedings of TSTT 2006, Third International Conference on Terminology, Standardization and Technology Transfer*, pages 265-272.

Vossen, P. (ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

Windhouwer, M.A. and Wright, S.E. (2012). Linking to linguistic data categories in ISOcat. In Chiarcos, C., Nordhoff, S., Hellmann, S. (eds), *Linked Data in Linguistics*, pages 99–107. Berlin: Springer.