FIN-CLARIN: A Framework for Depositing and Disseminating Language Resources for R&D

Atro Voutilainen and Krister Lindén
Department of Modern Languages
University of Helsinki
atro.voutilainen@helsinki.fi, krister.linden@helsinki.fi

Abstract

Researchers and developers in academia and industry would benefit from a facility that enables them to easily locate, licence and use the kind of empirical data they need for testing and refining their hypotheses and to deposit and disseminate their data e.g. to support replication and validation of reported scientific experiments. To answer these needs initially in Finland, there is an ongoing project at University of Helsinki and its collaborators to create a user-friendly web service for researchers and developers in Finland and other countries. In our talk, we describe ongoing work to create a palette of extensive but easily available Finnish language resources and technologies for the research community, including lexical resources, wordnets, morphologically tagged corpora, dependency syntactic treebanks and parsebanks, open-source finite state toolkits and libraries and language models to support text analysis and processing at customer site. Also first publicly available results are presented.

1 Introduction

Sharing of digital resources by and for researchers and other types of users is increasingly common worldwide, for instance there are several ongoing projects to create annotated text corpora and treebanks for various languages (Kromann, 2003; Mikulova et al., 2006; Nivre et al., 2006). In Finland, there are various kinds language resources for a number of languages at different organisations, but they are generally difficult to locate and take into use by researchers. Also their interoperability is generally poor due to lack of standardisation. There is an ongoing need for well organised, systematic and readily available language resources and tools. This paper outlines an ongoing effort to answer this need, in particular regarding the Finnish language.

We start with a description of language resources, users and their needs regarding language resources. Then we present an ongoing effort to answer these needs. Finally we outline some Finnish-language resources available currently or in the near future.

2 Resources, users and needs

2.1 Language resources

We use the term "language resource" to refer to a wide range of digital resources:

- small or large samples of naturally occurring text, speech and multimedia, representing different genres and time periods, and possibly annotated with various levels of linguistic analysis or other metadata;
- descriptions of the language (e.g. lexicons, morphologies, syntactic grammars, wordnets, ontologies) for human users;
- formal (linguistic or statistical) models of the language for automatic language processing tasks;
- tools to facilitate use of language resources;
- software and algorithms to enable automatic language processing tasks.

2.2 Types of users

Users of language resources are mainly researchers (in humanities; potentially also other fields such as computer and information science). Also commercial developers of language and information technological applications and services is a potentially large user segment, as development

of high-quality language technological solutions from scratch is a work and expertise intensive task.

2.3 User needs

Language resource users need means to identify and use interoperable language resources. The less effort the researchers and developers need in determining the existence of the required resource and in negotiating the access and use of the resource, the more time and money can be spent on research and innovation. Here is a partial "wish list" of user needs:

- researchers need empirical data to facilitate formulation, testing and evaluation of scientific generalisations;
- to enable replication of published empirical experiments, researchers need a way of sharing their empirical data, documentation and tools;
- researchers also need a facility for persistent storage and sharing of their (annotated) data (i) to help other researchers build on rather than duplicate existing work and (ii) to facilitate evaluation and recognition of an existing contribution, as discussed in (Pedersen, 2008);
- researchers need access to well-documented and modifiable language technological software to enable them to (i) annotate corpora specific to their research need and (ii) provide a "customised" annotation for a better match e.g. with the corpus linguistic research need;
- language technology companies and system integrators need access to well-documented and modifiable language technological software to help them provide a wider range of solutions and services to answer end-user needs in information discovery, multilingual communication, education, etc.

3 Solution in outline

FIN-CLARIN partners with Finnish service providers, research organisations, publishers and archives to set up the following kind of "ecosystem":

- a web service is set up at a service provider (Centre for Scientific Computing CSC) where language resources can be deposited, annotated and licensed for research and commercial uses;
- to help the user (researcher, developer) determine whether the service contains a relevant kind of language resource needed e.g. for formulation and testing of scientific hypotheses, the web service includes a workflow for metadata creation and use in combination with a search functionality;
- to help start use of the relevant resource, the web service sets up a transparent uniform licensing policy using which researchers can optimally access the resource as employee of web service member organisation on a single-access basis. In case the resource is not open source, licensing conditions can be understood easily on the basis of visual "laundry symbol" type classification (Oksanen et al., 2010);
- the service aims to offer various types of language corpora for researchers and developers: text, speech and video with varying levels of manually or automatically assigned linguistic annotation (e.g. morphological, syntactic, ontological). These corpora will represent both present-day Finnish (e.g. publicly available text collections on the Internet, e.g. Europarliament and Wikipedia texts) as well as diachronic corpora (licenced from domestic research institutions);
- in addition to extensive samples of natural language, the service also aims to provide various types of linguistic descriptions of the language, e.g. morphological lexicons, wordnets, name resources and grammatical descriptions (like valency descriptions). Such resources can be used for a variety of academic and practical purposes, e.g. reference material for linguistic studies, language learning solutions, and creation of language analysis software;
- to help researchers and developers efficiently use language corpora and linguistic descriptions, the service also offers a variety of software tools and technologies. One (frequent) type of researcher - a linguist with limited programming skills - needs user-friendly flexible tools to annotate, visualise

and quantitatively analyse the relevant corpus data available at the service (or even other corpora). – Another type of user is a researcher/developer with more extensive programming skills, who will benefit from a wider range of available open-source tools and technologies, e.g. software libraries and statistical modelling and analysis packages.

- the service aims to operate at a large scale, to offer very large quantities of language data (billions of words) to a growing number of users. FIN-CLARIN will partner with publishers, archives and other data providers to increase language resource coverage. FIN-CLARIN and its research partners conduct research to support annotation of the language data with an increasing level of informativeness and accuracy;

- users of the service sometimes enrich the data licenced from the service with additional annotation, e.g. as part of an empirical experiment reported in a scholarly publication. The service will offer a routine for such users to deposit their added annotations to the service for other users e.g. to enable validation and replication of empirical observations; different versions of the language data can be identified with persistent identifier codes (PIDs) and retrieved even a long time after their deposition (continuity of the service).

- the initial user base is expected to be mainly Finnish researchers and developers, but in the longer run the service aims to operate at European level (along with other CLARIN centres);

4 Current Offerings

In this section we outline some ongoing developments and resources available for FIN-CLARIN users.

4.1 FinnWordNet – the Finnish WordNet

FinnWordNet[1] is a lexical database for Finnish. It contains words (nouns, verbs, adjectives and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with relations such as hyponymy and antonymy, creating a semantic network. FinnWordNet can be used in language technology research and applications. It can also be used interactively as an electronic thesaurus. The first version of FinnWordNet has been created by having the words of the original English (Princeton) WordNet (version 3.0) translated into Finnish by professional translators.

4.2 FinnTreeBank – a Dependency Syntactic Treebank for Finnish

The FinnTreeBank project[2] is creating a manually annotated dependency syntactic treebank and an automatically created large parsebank for Finnish. This work is licensed under a GNU Lesser General Public License v3.0.

The first version of the treebank (Voutilainen et al., 2011) is annotated by hand and based on 19.000 example senctences in the Large Grammar of Finnish VISK - Iso Suomen Kielioppi (`http://kaino.kotus.fi/visk/etusivu.php`, (Hakulinen et al., 2004)). A parsebank for Finnish based on the Europarl corpus and the JRC-Aquis corpus will be published in late 2011.

4.3 Open Source Morphologies – OMor

The Helsinki Open Source Morphology Project for various languages aims at implementing full-fledged morphological analysers for a number of languages using the Helsinki Finite-State Transducer Technology (HFST).

The first large-scale implemented lexicon is an Open Source Finnish Morphology (OMorFi[3]), but a number of other analyzers and generators based on open source resources for various languages have also been implemented. These works are licensed under the GNU Lesser General Public License v3.0 unless specific restrictions apply to the original lexical resources for a language. The Finnish lexicon has been substantially extended and revised before it was compiled into a finite-state transducer, whereas the other languages are more or less mechanically derived from their repective sources.

[1] `http://www.ling.helsinki.fi/cgi-bin/finclarin/fiwn.cgi`

[2] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/`

[3] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/omor/index.shtml`

4.4 Helsinki Finite-State Transducer Technology (HFST)

The Helsinki Finite-State Transducer software[4] is intended for the implementation of morphological analysers and other tools which are based on weighted and unweighted finite-state transducer technology. This work is licensed under a GNU Lesser General Public License v3.0. The feasibility of the HFST toolkit is demonstrated by a full-fledged open source implementation of a Finnish lexicon as well as a number of other languages of varying morphological complexity (OMor) (Lindén et al., 2009).

Acknowledgments

The ongoing project has been funded via CLARIN, FIN-CLARIN, FIN-CLARIN-CONTENT and META-NORD by EU, University of Helsinki and the Academy of Finland.

References

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen and Irja Alho. 2004. *Iso suomen kielioppi* [Large Finnish Grammar]. Helsinki: Suomalaisen Kirjallisuuden Seura. Online version: http://scripta.kotus.fi/visk URN:ISBN:978-952-5446-35-7.

Matthias Kromann. 2003. The Danish Dependency Treebank and the underlying linguistic theory. *Proc. of the TLT 2003.*

Krister Lindén, Miikka Silfverberg and Tommi Pirinen. 2009. HFST Tools for Morphology – An Efficient Open-Source Package for Construction of Morphological Analyzers. *Proceedings of the Workshop on Systems and Frameworks for Computational Morphology 2009*, Zürich, Switzerland.

Marie Mikulova, Alevtina Bemova, Jan Hajic, Eva Hajicova, Jiri Havelka, Veronika Kolarova, Lucie Kucova, Marketa Lopatkova, Petr Pajas, Jarmila Panevova, Magda Razimova, Petr Sgall, Jan Stepanek, Zdenka Uresova, Katerina Vesela, and Zdenek Zabokrtsky. 2006. Annotation on the Tectogrammatical Level in the Prague Dependency Treebank. Annotation Manual. Technical Report 30, UFAL MFF UK, Prague, Czech Rep.

Joakim Nivre, Jens Nilsson and Johan Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006).*

Ville Oksanen, Krister Lindén and Hanna Westerlund. 2010. Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC2010).*

Ted Pedersen. 2008. Last Words: Empiricism Is Not a Matter of Faith. *Computational Linguistics, Volume 34, Number 3, September 2008.*

Atro Voutilainen, Krister Lindén and Tanja Purtonen (forthcoming). 2011. Designing a Dependency Representation and Grammar Definition Corpus for Finnish. *Proc. CILC 2011 - III Congreso Internacional de Lingüística de Corpus.*

[4] `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/hfst/index.shtml`