Towards a national infrastructure for terminology in the framework of the CLARA and CLARINO projects

Gisle Andersen

Marita Kristiansen

Norwegian School of Economics and Business Administration (NHH)

Summary

This paper outlines the plans for a national infrastructure for terminology in Norway. Such an infrastructure would constitute a knowledge base that makes substantial terminological resources available for users with various needs for quality-assured and structured scientific knowledge, via a single portal. The most important purpose of this effort is to enhance professional communication and understanding in a number of scholarly areas. There is a clear need for an electronically and openly accessible resource that provides harmonisation of conceptual understanding and term usage within defined scholarly areas across research institutions. This would not only benefit researchers and students, but also society at large. A national terminological infrastructure is aimed at meeting the challenges of the globalised research society and lay foundations for cooperative R&D work across academic and industrial institutions.

In the paper we first present briefly the background for such an initiative and the need for an infrastructure. Next, we present the projects CLARA and CLARINO, which can both be seen as providing important steps towards realising a national termbase. As the names suggest the two projects are closely related and are in fact both affiliated with the European CLARIN project. Finally, we draw some conclusions and present future perspectives.

1 Introduction

With its research focus on translation, language for specific purposes, terminology and business language, NHH's Department of Professional and Intercultural Communication has long been an important provider of terminology, for instance through specialised dictionaries on Norwegian-English terminology (Hansen & Lind 2010, Hansen & Lind 2008 and Lind 2007). In addition, the department has been a central player in advocating efforts and technology development for managing and accessing terminology (Brekke et al. 2006; Andersen 2008; Andersen and Kristiansen in press; Kristiansen 2010). The overall aim of the projects described in this paper is to provide and facilitate the use of a knowledge base by Norwegian research communities. Whereas a large number of terminological resources already exist in many European countries, e.g. the Swedish *Rikstermbanken*, Tilde's *EuroTermBank* and EU's multilingual termbase *IATE*, to mention a few, corresponding Norwegian resources are highly fragmented and lacking in common practices and standards on which the resources are built. The terminological coverage is also scarce, leaving many scholarly areas in danger of domain loss, not only in scientific discussions, but also in teaching and more popularised communication situations.

An early project co-funded by Unifob AKSIS (now Uni Digital) and the Norwegian Language Council has established a pilot version of a national terminological infrastructure called *Termportalen* (Andersen & Kristiansen in press). In this small-scale pilot project, problems concerning integration of disparate terminological resources into a common

searchable infrastructure were overcome, and a system for search across termbases was developed. However, a resource survey compiled in the project shows that existing terminological resources are considerably lacking in coverage and in need of updating, as stressed in the governmental White Paper *Mål og meining*. Also, the existing terminology resources are largely heterogeneous with respect to formats, content, structure and metadata. Two other projects, the CLARA project (Common Language Resources and their Applications) and CLARINO (Common Language Resources and Technology Infrastructure Norway), can be seen as providing important steps towards realising a national termbase.

The first project, CLARA, is funded by the EU's 7th Framework Programme and aims at developing knowledge and providing training for young researchers in the field of language resources and technology. NHH is one of nine partners in this network and is responsible for a subproject in CLARA, specified as a work package entitled *Harmonisation of Terminological Resources*. The members of CLARA are also participants in the European-wide CLARIN effort (Common Language Resources and Technology Infrastructure), enabling the coordination and sharing of cross-European knowledge resources. CLARIN was listed as a mature proposal on the ESFRI 2006 roadmap and received an EC grant for the preparatory phase which started in January 2008. It will run until the end of June 2011. Starting in July 2011, CLARIN will enter its construction and exploitation phase, an effort which will be organised as a so-called ERIC (European Research Infrastructure Consortium).

The second project, CLARINO is the name of the Norwegian branch of CLARIN. This is a planned project which will start in May 2011, if given the necessary funding from the Research Council of Norway. The aim of this project is to establish a Norwegian network of centres offering infrastructure services and to realise the Norwegian participation in CLARIN.

In both the CLARA and CLARINO projects, NHH takes a leading role in developing systems and standards for accessing, handling and disseminating terminology. In this paper, we describe the overall aims of these two projects and the interrelation between them.

2 The CLARA project

Being a Marie Curie Initial Training Network, CLARA aims primarily at offering research training to young researchers within language resources and technology. The scientific objectives of the CLARA research context is firstly to develop the next generation of data-intensive language models and applications by integrating approaches across language and country boundaries. Secondly, the project aims at contributing to the establishment of a pan-European infrastructure for language resources.

NHH's role in this project is primarily connected with one specific work package called *Harmonisation of Terminological Resources*. This subproject is a cooperation between four organisations, the Bergen-based institutions NHH and Uni Digital, Tilde (Latvia) and Temis (Germany). The objective of the project is to carry out research relating to the task of harmonising existing language resources. The project will have a special focus on terminology and other multilingual resources.

The integration of termbases and related data is a task which poses problems at both conceptual and technological levels. Although there are a large number of terminological

resources in Europe, many resources are fragmented or located indifferent institutions or in different formats. In CLARA the focus will be on developing language-independent methods and tools for constructing and consolidating multilingual termbases and ontologies. Furthermore, the project aims at developing methods and tool for corpus-based term extraction and for ontology-based domain recognition of text. Also, we propose to investigate the degree of compatibility of different termbases at the top level of domain classification, at the mid-level of sub-domain, and at the lowest conceptual level, i.e. the concept itself and its relations to neighbouring concepts within the sub-domain.

Integration of termbases is a notoriously challenging task, as was evidenced by several contributions in a research seminar organised in Bergen this autumn (Kristiansen 2010; Nilsson 2010; Schumann 2010; Vasiljevs 2010; Warburton 2010). Seemingly simple questions such as how to register terms (main terms versus its synonyms) are not necessarily answered the same way in all termbases. Also, how to handle specialised lexical collocations in language resources as parallel corpora, specialised dictionaries and term bases are something which needs further research, and a PhD project, Patiño (2010), is devoted specifically to this topic and investigates corpus data based on Free Trade Agreements. Moreover, the task of domain classification is very often left unresolved, at least at the more detailed level (Kristiansen 2010). The CLARA project will therefore attempt to develop new, innovative approaches to cross-termbase integration which should handle these questions.

3 The CLARINO project

Representing the Norwegian branch of CLARIN, the CLARINO project aims to provide a home for a wide range of existing language resources and adapt them to the standards developed in CLARIN. Among the scientific goals of CLARINO is to develop a centre of expertise in the creation and development of terminological language resources and technologies to enhance professional communication in Norwegian through the development and dissemination of terminology. It is in this task NHH has a central role through a work package referred to as CLARINTerm.

The main deliverable of CLARINTerm is a technical architecture that integrates structured, mono- and multilingual knowledge bases via a common web-based portal. The terminology portal will give public and controlled unified access to distributed terminology resources in a grid-based umbrella architecture. This architecture will then be linked to European terminology resources via CLARIN. We plan to offer both distributed and centralised solutions to providers and developers of terminology. Thus, CLARINTerm will cater for autonomous bases with a common interface and search system, as well as integrate a variety of existing termbases into a central termbank. The termbase will contain entries for concepts and conceptual descriptions such as main terms and synonyms, definitions and concept relations. It will be the decision of each data provider to either have the data fully integrated into the central termbank or made accessible as a distributed base via the CLARINTerm portal. Moreover, hybrid solutions may be offered, in which an instantiation of CLARINTerm draws from both a local source and the central termbase, but is presented uniformly as one resource.

Representing a terminology stronghold in Norway, the CLARINTerm research group and development team from NHH, the University of Bergen (UiB) and Uni Digital, respectively, wants to establish a national infrastructure for terminological language resources and

technologies. This is a response to the national responsibility now given to higher education (cf. Act relating to Norwegian higher education § 1-7) to develop and disseminate Norwegian terminology. This will be a natural extension of already existing resources originating from the Norwegian Term Bank which in recent years have been further developed in projects such as KB-N (Brekke et al. 2006), *Mikroøkonomen* (Kristiansen 2010), *Termportalen* (Andersen & Kristiansen, in press) and CLARA (https://clara.uib.no/). At present no such national infrastructure exists, and establishing CLARINTerm will therefore be an important contribution to the harmonisation of Norwegian terminology.

Moreover, the project will offer advisory services and establish a best practise for the harmonisation and expansion of mono- and multilingual concept-based termbases, including general methodological aspects, standards for annotation of terminological data and for presenting the data. This includes definition writing, handling of synonymy and duplicates, and technological aspects such as the transfer of termbase data (via metadata standards such as the TBX exchange format). CLARINTerm will also be a developer of and a repository for different computational tools relating to terminology management. This may include tools for distributed management of termbase entries, for efficient data conversion, for organisation and graphic display of ontology relations, for text-based extraction of terminology, for linking termbase entries to relevant domain-specific text corpora, etc. The centre will deplore and disseminate guidelines and standards, dissemination and administration of terminological language resources.

4 Who needs a national terminological infrastructure?

A national terminology infrastructure will be an effective instrument in dealing with the global and national knowledge challenges facing our society, and, we argue, lead to increased competitiveness for the Norwegian research community. Out hope is that it will be an important tool for both teaching and research training at university level, as well as for the subject specialists' ability to communicate efficiently within and beyond their expert groups.

The proposed activities of the national infrastructure are largely concurrent with the strategic priorities stated in *Verktøy for forskning*[1] and the thematic and technological priorities of *Vilje til forskning*[2]. Observing the aims of these strategic documents, the project will lower the threshold of internationalisation of research, as the resources to be developed in this infrastructure are concept-based and multilingual. It will also contribute to bridging the communicative gap between experts and the public, as the resources will provide extensive conceptual information, definitions and graphic displays, as well as web-based interactive facilities.

We believe that such a project will also contribute to cost-efficiency at several levels; the user can retrieve information from a wide range of relevant resources via a single portal and is guaranteed the quality of the retrieved information using this national knowledge base. The knowledge base will consist of data that is continually updated. Importantly, the users of the infrastructure will have access to ontologically structured domain-specific knowledge enabling communication and learning in the user's mother tongue.

[1] Verktøy for forsking – Nasjonal strategi for forskningsinfrastruktur (2008–2017), RCN.
[2] St.meld. nr. 20 (2004–2005)

It is important that a national terminological infrastructure is established with the view to cater for the needs of different kinds of user groups – in close cooperation with these user groups. This requires that the language resource and technology infrastructure is comprehensive, versatile and flexible, and serving the need of a wide range of users. These include the *researcher*, who needs to discuss theories and concepts with colleagues or wishes to streamline the content of a research paper with the officially recognised terminology, the *academic scholar*, who wishes to prepare his lecture notes applying a terminology which is based on a broad consensus between a nation-wide network of colleagues within his scholarly domain, the *translator*, who needs to know the official term in connection with translations for a public customer, the *student*, who needs to look up the domestic term when preparing an assignment based on English literature, the *journalist*, who needs to know what domestic terms exist for a new product that is the topic of his next article, the *business developer*, who wishes to extend his services to a new field, basing the business plan on officially recognised terminology of the field in question, as well as a wide array of *other user groups*. A national terminological infrastructure will also offer a means to cross-disciplinary or cross-institutional studies, in which conceptual convergence is vital to establish a common ground for scientific discussions.

These objectives are crucial to the terminology branch of CLARINO, which will offer terminology-related services to a variety of user groups. The project will enable developers of terminological language resources, such as the Norwegian Foreign Secretary, to disseminate their resources and have the terminology data made accessible nationally and internationally via the national CLARINO infrastructure and thereby also the international CLARIN infrastructure.

The proposed system will also allow owners of term lists and termbases data to upload their data in a simple format (e.g. tab separated file) and have it converted via available tools to exchangeable formats and in accordance with existing industrial standards. Termbase technology developers, such as Standards Norway, will have the opportunity to align their technology in accordance with national and international recommendations and best practises. We believe this effort can also be valuable for enterprises and academic institutions developing language technology systems for knowledge management and for checking of terminological consistency, multilingual technologies such as machine translation, etc. The ambition is to provide the technical solutions needed for the coordination, integration and further development of terminology nation-wide. This is seen as relevant for national bodies like the Norwegian Language Council and the Norwegian Association of Higher Education Institutions (UHR), who will be able to realise their stated goals of a coordinated national effort in terminology and termbase development, as recently expressed in governmental white papers like *Mål og Meining*. Moreover, the infrastructure is needed for what is also a strategic objective of academic institutions like NHH and UiB, who have recently established their institutional language policy documents advocating the use of Norwegian and parallel terminology. Effectively, the output of the infrastructure development will constitute the terminology component of the Norwegian Language Bank. Finally, for end users of terminology – from field experts to the general public – including translators and interpreters, the project will provide unified access to a wide range of updated terminological data; hence eliminate the need to check a variety of web locations in the search for specialist field knowledge.

5 What should a national terminological infrastructure look like?

A national infrastructure is a complex whole of technological and linguistic resources, consisting of at least four parts, i.e., a *system*, *methods*, *content* and *tools*. First, the *system* includes the integration of multilingual knowledge bases (termbases) via a common web-based portal with public and controlled access. Second, it is necessary to establish a common *method*, i.e., a consensus-based theoretical and methodological framework for conceptual structuring, including aspects such as definition writing and delineation of domains and subdomains. The third part is the *content*, which will involve a set of comprehensive and updated national termbases, containing multilingual entries for concepts and conceptual descriptions such as main terms and synonyms, definitions and concept relations, based on consensus between subject specialists in the relevant domains. Finally, existing *tools* for distributed management of termbase entries, for efficient data conversion, for hierarchical organisation and graphic display of ontology relations must be harmonised or further developed. Tools will also include applications for text-based extraction of terminology and for linking termbase entries to relevant domain-specific text corpora.

Fortunately, we are not starting from scratch, since the infrastructure can build directly on technology and resources available through the CLARINO network, as well as termbase technology developed over several decades in Norwegian projects mentioned in section 3 above. It will function both as a repository for existing terminological language resources, including "historical" resources that are no longer being updated, as well as giving direct access to current and continuously updated terminology. Naturally, users will be able to restrict search and access according to the various sources in the database.

The majority of the existing resources contain Norwegian and English terminology, and there is an immediate need for the Norwegian research community to further develop such bilingual terminologies. However, since the infrastructure will be concept-based, terminology in different languages can easily be added for areas where this is relevant. This makes it also possible to further develop cooperation with terminology providers internationally, beyond what is already planned in the CLARA project. The plan is to make this conceptual data both browseable and searchable. Domain and subdomain relations will be accessible to the user, who may search unrestrictedly, in multiple domains, or in a specific domain (including or excluding its subdomains) by ticking off relevant boxes. Linking and navigation between concepts is done via metadata representations of the semantic relations synonymy, superordinate, subordinate and coordinate concepts. The system will offer graphic representations of ontological structure and concept hierarchies, and visual semantic networks will be generated on-the-fly. Termbase entries will also include possible non-preferred synonymous terms that are issued with a user warning.

The system will provide access to terminology resources stored in a variety of relational databases and present them in a unified way by generating a well-defined XML representation of the data that can be flexibly presented in human-readable formats via XSLT style sheets via HTML and Javascript. The termbase system will be based on the current solutions used in NHH's existing termbase representing domains like economics and business administration. The current system is based on non-proprietary software, using a relational database (MySQL/PostGreSQL), but also uses in-house software based on LISP. Furthermore, it will interact with CLARIN's system for user control and be fully integrated with its accessibility system, distinguishing superusers from administrators and users with reading and edit accessibility to local termbases. The system will also offer full flexibility

with respect to web browsers and wiki functionality for uploading and editing of multimedia content, such as figures and videos, a functionality which is increasingly relevant for termbase entries. Finally, the resources made accessible via CLARINTerm will be adapted to and interoperable with the CLARIN and ISO standards, such as the TBX terminology exchange format.

The knowledge base terminologies are intended to cover a broad range of central concepts within each domain and to be based on consensus between subject specialists in the relevant domains. Entries in the termbases contain substantive multi-layered conceptual information and metadata descriptions that enable users to access data via any conceptual representation, including terms, definitions, concept relations, other semiotic representations, usage examples and potentially also a domain-specific text corpus, dependent on availability. Individual entries should minimally have an officially recognised term in Norwegian bokmål and/or nynorsk, a definition and the corresponding English term, but typically the entry will also contain other types of (meta-)information.

6 Methodological aspects

A national termbase is of course a large-scale effort that requires efficient systems for knowledge management and standardised modes of representation. There is a need for a consensus-based theoretical and methodological framework which can be instructional for content providers. Such a methodological basis should include descriptions of conceptual structuring, definition writing, delineation of domains and subdomains, applications of metadata, and various other topics that will have bearings on the practical work of the different terminology developers who wish to provide content to the termbank. A source of systematisation will be the standards and recommendations provided by ISO 704 (Terminology work – Principles and Methods) and 1087 (Terminology Work – Vocabulary) and also the framework established by CLARIN, where the project initiators contribute in the working group that develops standards and taxonomies for language resources.

As already mentioned, existing resources are largely heterogeneous. Thus the management of such resources requires efficient and flexible conversion tools. In addition, CLRAINTerm will utilise a wide array of more general computational tools that may speed up the process of creating new terminology in domains still to be developed. This includes tools for distributed management of termbase entries, for hierarchical organisation and graphic display of ontology relations, for corpus-based extraction of terminology and for linking termbase entries to relevant domain-specific corpora. Also a number of relevant language technological tools are made available in other branches of the CLARINO project, including taggers, parsers, tools for morphosyntactic annotation, treebanking tools, etc. For example, tools developed in an earlier RCN-financed infrastructure project, the Norwegian Newspaper Corpus, will be a useful resource supplement to generate relevant and topical specialist concepts. This project has developed systems for semi-automatic domain classification of text and for identification of collocations and multiword expressions, tools which are jointly valuable and reusable to determine termhood and extracting terminology from a wider set of text resources than strictly domain-specific texts. Thus we are advocating an eclectic approach of exploiting whatever resources are available, but a stringent approach with respect to what gets included in the quality-assured knowledge base.

7 Conclusion

In this paper we have argued in favour of a coordination of terminological resources at national and international levels. The effort we are proposing is to develop a national infrastructure for terminology. Such an infrastructure will provide a substantial, national and official knowledge base aimed at enhancing professional communication in a number of scholarly areas and at all levels of sophistication. The effort will provide a national delineation of scholarly areas, harmonisation and transparency of conceptual understanding and term usage within and across disciplines and research communities. It will also entail a construction and consolidation of domain-specific ontology, which will form a basis for conceptual discussions not only within the respective domains, but also across interrelated disciplines in cross-disciplinary projects. The facility will therefore be an important tool for professional communication for the benefit of researchers, students and the society at large. Such communication takes place not only internally between researchers within the same domains but also at a cross-disciplinary level, in teaching situations and research training at university level, and between subject specialists and society in general. Different terminologies will be needed in these situations, something which will be facilitated by the projects discussed in this paper.

References

Andersen, Gisle (2008) Terminologi som språkressurs og forskningsinfrastruktur. *NORDTERM* 15. 53-58.

Andersen, Gisle/Kristiansen, Marita (in press) Terminor og Termportalen – nye initiativer for norsk terminologisk infrastruktur. To appear in: *Proceedings from Nordterm* 16.

Brekke, Magnar/Innselset, Kai/Kristiansen, Marita & Øvsthus, Kari (2006) KB-N: Automatic Term Extraction from Knowledge Bank of Economics. *LREC 2006 - 5th International Conference on Language Resources and Evaluation*. 1912-1916.

Hansen, Einar/Lind, Åge (2010) *Norsk-engelsk økonomisk-juridisk ordbok.* Bergen: Cappelen Akademisk forlag.

Hansen, Einar/Lind, Åge (2008) *Engelsk-norsk økonomisk-juridisk ordbok.* Bergen: Cappelen Akademisk forlag.

ISO 704:2000, *Terminology work – Principles and methods*.

ISO 1087–1: 2000, *Terminology work – Vocabulary – Part 1: Theory and application*.

Kristiansen, Marita (2010a) Representing interrelated domains in termbases. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.

Kristiansen, Marita (2010b) Language Planning in Higher Education. The Case of Microeconomics. In Heine, Carmen/Engberg, Jan (eds) *Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009*. Aarhus 2010, http://www.asb.dk/fileexplorer/fetchfile.aspx?file=19126.

Lind, Åge (2007) *Engelsk-norsk juridisk ordbok. Sivil- og strafferett.* Bergen: Cappelen Akademisk forlag.

Nilsson, Henrik (2010) Rikstermbanken / Term-O-Stat – Experience from national term bank projects in Sweden. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.

Patiño, Pedro (2010) A corpus-driven study of specialized collocations in Free Trade Agreements. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.

Schumann, Ann-Kathrin (2010) Goals and Challenges in EuroTermbank. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.

St.meld.nr. 35 (2007-2008) *Mål og meining. Ein heilskapleg norsk språkpolitikk*. URL http://www.regjeringen.no/nb/dep/kkd/dok/regpubl/stmeld/2007-2008/stmeld-nr-35-2007-2008-.html?id=519923

Vasiljevs, Andrejs (2010) Consolidation of multilingual terminology resources. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.

Verktøy for forskning. URL http://www.roadmaptgi.fr/Documents/Norwegian%20strategy.pdf

Warburton, Kara (2010) Lexicographical and terminological resources: data models, relationships, applications. Paper presented at conference Terminology and Resource Harmonisation, Bergen, 13-17 September 2010.