ELSST: a broad-based Multilingual Thesaurus for the Social Sciences

Lorna Balkan*#, Ken Miller*, Birgit Austin*, Anne Etheridge*, Myriam Garcia Bernabé*, Pam Miller*

*UK Data Archive,
University of Essex,
Colchester,
CO4 3SQ
{balka, millk, birgit, aether, garcmn, millp}@essex.ac.uk

#Department of Language and Linguistics,
University of Essex,
Colchester,
Essex,
CO4 3SQ

Abstract

This paper describes the motivation for, and methodology behind the creation of ELSST (European Language Social Science Thesaurus), a broad-based multilingual thesaurus for the social sciences. The thesaurus was produced by the UK Data Archive (UKDA) as part of the EU-funded LIMBER (Language Independent Metadata Browsing of European Resources) project and was derived from their in-house English monolingual thesaurus, HASSET (Humanities and Social Science Electronic Thesaurus). The multilingual thesaurus is currently available in four languages English, French, German and Spanish and in various formats, including RDF (Resource Description Framework).

1. Introduction

This paper describes the motivation for, and methodology behind the creation of ELSST (European Language Social Science Thesaurus), a broad-based multilingual thesaurus for the social sciences. The paper is in the following parts. Section 2 describes the background and motivation for the thesaurus. Section 3 describes the purpose and structure of a thesaurus. Section 4 describes the two-stage construction of the thesaurus, namely the reduction and translation processes. Section 5 reports on how the thesaurus was evaluated. Section 6 discusses some of the future plans for the thesaurus and gives contact details.

2. Background

The goal of the LIMBER project (see reference: LIMBER project) was to facilitate cross-European data analysis independent of domain, resource, language and vocabulary. It sought to address the problems of linguistic and discipline boundaries, which, within a more integrated European environment, were becoming increasingly important. Decision-makers, researchers and journalists needed to be provided with a broader, comparative picture of society across the continent; with the social science information often required to be correlated with information from domains such as environmental science, geography and health. This cross-discipline interoperability would be provided via a uniform metadata description. In addition, the provision of multilingual user interfaces and the controlled vocabulary of a multilingual thesaurus (ELSST) would make these datasets globally accessible in a range of end user natural languages.

Central to the goals of LIMBER was the adoption of RDF (Resource Description Framework)(see reference: RDF) to allow interoperability between metadata standards and mappings between thesauri. RDF is a general framework for describing metadata about Web accessible resources developed by the World Wide Web Consortium (W3C). This framework is intended to provide a simple model for user communities to define their own metadata descriptions, which can then be interpreted throughout the Web, especially via automated processors. RDF is based on a simple graph model capturing the resources on the Web and the relationships between them, in a flexible and extensible manner. This is realised using an XML format which is compatible with other XML developments and tools. Thus by using RDF each user community can describe the properties of it own resources and combine them with the descriptions from other communities in a uniform manner. The LIMBER project also developed a RDF schema for thesauri so that these ontologies could be considered as metadata resources. Thus RDF would allow the use of metadata across domains and mappings between thesauri.

3. What is a thesaurus?

A thesaurus is a list of controlled vocabulary, or keywords, displayed in a hierarchical structure. Its purpose is to facilitate the user, both the indexer and the searcher, in choosing the most suitable terms and to use terms consistently.

The arrangement of terms into a hierarchical structure helps a user to broaden a search or make it more specific. Synonyms, non-preferred (lead-in) terms, and related terms provide the user with suggestions for further useful search terms.

In general a thesaurus helps users to define their search in the terms which are most likely to lead to retrieval of relevant information. Retrieval failure is inefficient and can be costly. A thesaurus can help us to more fully exploit our stores of information and reap from past and continuing investment.

In the context of a distributed information retrieval system, a common indexing and retrieval tool is essential in order to maintain effective and balanced retrieval across individual systems and enable a user to search the standardised metadata common to all of the sites. By providing a multilingual thesaurus as a retrieval tool, a user may search in his preferred language but still be guaranteed all of the advantages of the effectiveness and consistency of a thesaurus. Users can formulate queries in their preferred languages and retrieve all relevant (meta-) data objects in whatever language the objects are stored.

ELSST is based on thesaural relationships and displays as outlined in ISO 2788-1986 guidelines for the establishment and development of monolingual thesauri (see ISO 2788-1986, (1986)), and ISO 5964:1985 guidelines for the establishment and development of multilingual thesauri (see ISO 5964:1985, (1985)). It employs the conventional range of term relationships of equivalence (preferred and non-preferred terms, USE/UF), the hierarchical relationships (broader and narrower terms, BT/NT) and the associative relationship (related, RT). Additionally, translational equivalences are defined between terms in the four different languages of the thesaurus: English, French, German and Spanish (see Section 4.2).

4. Methodology

Construction of the thesaurus proceeded in two stages: first the monolingual thesaurus was reduced, and then the translation of the reduced thesaurus was carried out. We discuss each stage in further detail below.

4.1. Reduction of monolingual thesaurus

The English monolingual thesaurus was created from the existing thesaurus of the UK Data Archive (HASSET[1]), which had been developed over 25 years to index the holdings of the UKDA, and hence considered the definitive thesaurus for social science data archives. It has approximately 4,200 preferred terms in 250 hierarchies.

The following policies were adopted for the task of restructuring the UKDA HASSET for use across Europe. A top down reduction to a very broad base was made, removing any cultural or institutional bias, of the major 26 hierarchies. These reduced and restructured listings were circulated to the user group, which consisted of the CESSDA archives (see reference: CESSDA) to determine whether they still met their needs. However, the reduction was made with the understanding that the rationale behind ELSST was to produce a common ontology which could be extended via local extensions to cater for the cultural and institutional needs of the individual archives and also allow for inclusion, via mappings, to specialised thesauri in certain subject areas.

The 26 major hierarchies were selected by a review of the most frequent HASSET terms used as keywords in the UKDA catalogue. This resulted in an initial 10 hierarchies, the terms of which constituted nearly half of the complete thesaurus. The RT relationships of the terms of these hierarchies were then examined to determine which additional hierarchies to add. 16 further hierarchies were then added which had terms with the greatest number of associations with the terms of the original 10 hierarchies. This resulted in approximately 2,500 preferred terms prior to reduction.

The following is a list of the hierarchies initially selected for ELSST:- Economics, Labour and Employment, Politics, Political Systems, Social Problems, Discrimination, Attitudes, Disadvantaged Groups, Political Institutions, Ethnic Groups, Living Conditions, Social Structure, Data, Age Groups, Demography, Sociology, Social Welfare, Environmental Sciences, Education, Identity, Nationality, Families, Religion, Analysis, Methodology and Family Environment.

The process of removing any cultural or institutional bias resulted in a much greater reduction than expected, with the initial version of the monolingual ELSST containing just under 1,000 preferred terms. However, feedback from CESSDA, both to the monolingual structure and during the translation process, along with the feedback from several workshops led to the addition of a further 23 associated hierarchies and a final published monolingual version (30/5/2001), which had 49 hierarchies and 1,380 preferred terms.

4.2. Translation of ELSST

ISO 5964 recognizes three approaches to the construction of multilingual thesauri:

1. Ab initio construction: i.e. the establishment of a new multilingual vocabulary without direct reference to the terms or structure of an existing thesaurus;
2. Translation of an existing monolingual thesaurus
3. Reconciliation and merging of existing thesauri in two or more working languages.

Examples of (1) include the cultural heritage thesaurus HEREIN (see reference: HEREIN thesaurus). An advantage of this approach is that it is easier to ensure language neutrality (i.e. lack of bias towards any one language). However, the costs of producing such a thesaurus are considerable, since the structure of the thesaurus has to be established, as well as the definition and multilingual equivalence of its terms.

An example of (3) is the mapping between the AAT (Art and Architecture Thesaurus), the thesaurus of the Royal Commission of Historical Monuments of

[1] HASSET was initially based on the 1977 UNESCO Thesaurus, ISBN 92-3-101469-2.

England (RCHME) and the French Mérimée thesaurus (see reference: Mérimée thesaurus). There is currently a lot of interest in the mapping between thesauri (see Doerr, 2001). Problems include the difference in the hierarchical structure of the relevant thesauri, as well as differences in the semantics of the actual terms. While equivalency is sought between terms, this does not imply that the hierarchical structures themselves must also be equivalent.

The ELSST thesaurus was constructed by method (2). The definition of "translation" can be differently understood, depending on the intended use of the resulting translation. On the one hand, translations may be used principally to allow the users of the target languages to better understand and use the source language terms. In this case, the translations may not be suitable indexing terms in the target language. If, on the other hand, it is intended to use the target terms as indexing terms in their own right, they may not correspond to the best translation of the source terms. ELSST adopts the second strategy, making the translation process more akin to method (3) than to conventional translation tasks.

ISO 5964-1985 (ibid.) guidelines stipulate that, regardless of the method of construction, the languages of the resulting multilingual vocabulary must have equal status to the source language or languages. This is also the goal of ELSST. In addition, ELSST strives to be multicultural. The term multicultural is intended to mean that preferred terms should reflect a European rather than a national or local approach. However, terms that are specific to a language or region are permitted as non-preferred terms.

As with method (3) above, the notion of equivalence in ELSST applies to terms only, and not to hierarchical structures. Thus the different language versions of the thesaurus may have different hierarchical structures, although in version 1 of ELSST structures are identical. As far as term equivalence is concerned, only preferred terms require a translation equivalent, since it is not always possible or appropriate to find a translation of a non-preferred term (in the case of spelling variants or language specific terms, for example).

ISO 5964-1985 (ibid.) defines a classification scheme for different types of term equivalence: exact equivalence, partial equivalence, single-to-multiple equivalence, inexact equivalence and non-equivalence. These equivalence relations are being used with increasing frequency in thesaurus mappings: in the HEREIN project (ibid.), the Mérimée project (ibid.), and other projects. The taxonomy of equivalence relations was used in the construction of ELSST to help identify different types of translation problems and possible solutions (see Section 4.2.1 below). In future versions of ELSST we may adopt a formal representation of these different types of relation. Since the standard is in the process of being revised, we await the outcome of the revision process.

A novel solution to the problem of translation mismatch between terms is the use of a translation scope notes in ELSST. These are used to explain translation mismatches between the English term and a translation equivalent. (In future versions of ELSST their use could be extended to explain translation mismatches between other language pairs.)

Translation was carried out by a team of translators at the UKDA, who met on a regular basis to discuss problems as they arose. They provided feedback to those working on the monolingual thesaurus, so that changes to the monolingual thesaurus, such as the addition of scope notes, could be implemented where necessary. Verification of the translations was carried out by bilingual information experts at the appropriate CESSDA sites.

As sources for translations, other thesauri in the same domain, such as the ILO (see reference: ILO thesaurus) and the UNESCO thesaurus (see reference: UNESCO thesaurus) were consulted where possible. However, other thesauri had to be used with caution, since terms were not always used in the same sense as in ELSST. Thus "DRUGS", which includes both legal and illegal drugs in ELSST is translated as "MEDICAMENTO" (medicinal drug) in the UNESCO thesaurus (ibid.), while it is translated as "DROGA" (illegal drug) in the ILO thesaurus (ibid.), neither of which is appropriate as a translation in ELSST. (We discuss the translation we adopted for this term in ELSST in Section 4.2.1 below.)

4.2.1. Translation problems and some solutions

There were a number of terms found that have no direct equivalent in one or more of the target languages. This was due to a variety of reasons. Firstly, despite our best efforts in the reduction of HASSET, some culture-specific terms made their way into ELSST. Examples include the term "PRIME MINISTER", which translates literally as "PREMIERMINISTER" in German but which corresponds to "BUNDESKANZLER" ("chancellor" in English). The solution here was to adopt the more general term "HEADS OF GOVERNMENT" as the preferred term and retain the culture-specific terms as non-preferred terms.

In some cases, a morphological mismatch between the languages accounted for missing terms. For example, "non" in "NON-PROFESSIONAL OCCUPATIONS" is used in a neutral sense, while the corresponding prefix in German "nicht", always expresses an opposite. Thus "nicht professionell" in German corresponds to "unprofessional" rather than to "non-professional".

In other cases it is less obvious why concepts are not lexicalised in the target language. For example there is no direct equivalent term for "HOMELESSNESS" in Spanish. The solution was to translate it by a closely related term, "DESAMPARADOS" (literally "the homeless") in Spanish.

Sometimes the meaning of a source language term corresponds to the meaning of more than one term in the target language (*single to multiple equivalence* in ISO terminology). For example, as noted above, "DRUGS" in English encompasses both legal and illegal drugs. In all the other three languages, there is no collective term for these two types of drugs. Thus in order to translate "DRUGS" in ELSST a synthetic term consisting of both types of drugs had to be created. The German for

"DRUGS" thus became "DROGEN UND MEDIKAMENTE" (literally "illegal drugs and medicinal drugs").[2]

In other cases a source language term may express either a broader or narrower concept than its target language equivalent (*partial equivalence* in ISO terminology). An example where the source language term is broader in meaning than its target language equivalent is "VOCATIONAL EDUCATION" in English and its German equivalent "BERUFSBILDUNG". ("VOCATIONAL EDUCATION" includes both the educational side of vocational education as well as the practical side, while "BERUFSBILDUNG" is predominantly education-based.) Here, a translation scope note is used to explain the difference in concept. Where the source language term expresses a narrower concept than its target language equivalent a qualifier was frequently used to narrow the meaning of the target language term. This was obligatory where more than one source term shared the same translation. An example is "secuestro" in Spanish, which can mean both "hi-jacking" and "kidnapping". To distinguish the two translations, "HI-JACKING" was translated as "SECUESTRO (VEHICULOS)" and "KIDNAPPING" as "SECUESTRO (PERSONAS)". Sometimes, however, it was not possible to find suitable translations for closely related terms in the source language that share the same translation in the target language. An example is "VOCATIONAL TRAINING", "JOB TRAINING" "OCCUPATIONAL TRAINING", and "PROFESSIONAL TRAINING" all of which translate as "FORMATION PROFESSIONNELLE" in French. The decision was taken to make the broadest term (i.e. "OCCUPATIONAL TRAINING") the preferred term, and let the other terms function as non-preferred terms in English.

Often the problem was not one of missing concepts but of partially overlapping concepts in the source and target languages (*inexact equivalence* in ISO terminology). An example is "PROFESSIONAL OCCUPATIONS" which translates as "PROFESSIONS LIBERALES" in French, but which is not totally synonymous with it (the set of occupations belonging to "PROFESSIONAL OCCUPATIONS" only partially overlaps with that belonging to PROFESSIONS LIBERALES"). Again the solution was to explain the differences in a translation scope note.

Sometimes the source language term has a corresponding target language translation, but they have different connotations. An example is "INNER CITIES", which is associated with social deprivation in English. In France, social deprivation is associated more with the suburbs than with inner city areas, thus a straight translation of the term fails to capture the appropriate connotation. The solution in this case was to use the English expression as a loanword translation in French.

Even where a translation of a source language term was available in the target language, it was not always used, since some other term sounded more natural as an indexing term in the target language. This was the case with the English term "SPEECH DEFECTIVE" where the French equivalent "HANDICAP DE LA PAROLE" (literally "speech disability") was chosen over a more literal translation.

Most mismatches were between terms. We did, however, find an interesting example of mismatch at the structural level. "UNDER-AGE DRINKING" is a non-preferred term of "DRINKING OFFENCES" in the English thesaurus, but this is inappropriate in the French thesaurus, since under-age drinking is not an offence in France.

A concern during the translation process was to avoid introducing ambiguity in the target language where none exists in the source language. It sometimes happened that homonyms (i.e. words that are written the same but are unrelated in meaning) were the translation of two different source language terms. For example, both "AGE" and "OLD AGE" translate as "Alter" in German. In order to distinguish them, qualifiers were added, so that "AGE" translated as "ALTER (ALLGEMEIN)" and "OLD AGE" translates as "ALTER (RENTENALTER)". Qualifiers were added even where homonyms were not involved in order to disambiguate a target language term that was felt to be potentially confusing. An example is "COMPANIES" which translates as "SOCIETES (ECONOMIE)" in French.

5. Evaluation

The Council of European Social Science Data Archives (CESSDA) promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe. CESSDA members were consulted at each stage of the development of ELSST. A requirements workshop was organised at the beginning of the project, to which CESSDA members and other information and social science specialists were invited. This resulted in a number of recommendations about the form and content of the thesaurus, which were taken into consideration during its construction. Mechanisms for appraisal were set up and worked very well. However feedback from these mechanisms continued way beyond the planned cut off date.

CESSDA members were sent the ELSST monolingual hierarchies as they were produced and were asked to comment on the structure of the hierarchies, as well as on the individual terms. Feedback was used to produce a pre-final version of the monolingual thesaurus, which was then evaluated at a further workshop held in conjunction with the Conference of the International Association of Social Science Information Science Technologies (IASSIST) in May 2001 in Amsterdam. Members of the wider international community also attended this workshop. Participants were asked to complete a series of exercises which demonstrated different ways of viewing and using the thesaurus either for indexing or searching for data. The thesaurus was linked to the UKDA database of datasets, so that searches, using ELSST, could be performed in a realistic situation. Participants were asked to suggest missing terms and make any other comment on the hierarchies.

[2] ISO guidelines recommend that cases such as these are translated using combined terms (e.g. "DROGEN + "MEDIKAMENTE"). We are currently investigating the use of combined terms, and the syntax and semantics or their combine operators.

Some of the translation work was produced in tandem with the monolingual hierarchies and sent to the appropriate CESSDA members for comment. Most of the translation was not, however, assessed until the end-of-project workshop which was held in September 2001 at the University of Essex in the UK for the User Group and wider European community. Evaluation took the form of an exercise, which built upon that used at the IASSIST workshop but which was now available in the four different languages of the project. Additionally, users were asked to review one or more hierarchies of their choice and comment on their structure, content, or translation. A questionnaire was completed at the end. Reaction to the thesaurus was generally favourable. Nearly 70% of users thought the hierarchies were well structured. The majority of participants said they would find the thesaurus useful for indexing/retrieval purposes, and found the scope notes useful. Very few had negative comments about any aspects of the thesaurus.

Given the difficulty of evaluating the thesaurus adequately within the short time frame of a workshop, CESSDA members were also sent the pre-final multilingual thesaurus mounted on a database to review at their leisure. Very extensive feedback was received from CESSDA members at this stage, in the form of additional non-preferred terms, alternative translations, and other comments. The final version of the thesaurus was produced based on feedback from CESSDA members and from the workshops.

Indexing work using ELSST is currently being carried out at the UKDA, and results from this will feed into the next version of ELSST.

6. Conclusion

Prior to developing ELSST only a few CESSDA members had adopted thesauri for their resource discovery systems. The evaluation feedback from CESSDA members and other organisations was very positive in terms of adapting both the scope and depth of the ELSST thesaurus and in making the multilingual thesaurus a valuable tool in locating and interpreting resources that can be used for comparative research.

The success of the LIMBER project is perhaps best reflected in the further initiatives and projects that will continue the achievements of the project.

The CESSDA members wish to adopt the ELSST thesaurus as the controlled vocabulary for their virtual catalogue of European data resources, and the Social Science Data Archive of the Netherlands have agreed to allow use of their Thesaurus of Social Research Methodology (SRM) in the same catalogue.

Further languages of Finnish, Norwegian, Danish and Greek are proposed in a future EU project and the RDF schema for thesauri developed in the LIMBER project will be taken forward and proposed as an international standard for the interchange of thesauri.

However, further work is required on the thesaurus. The methodology listing from the thesaurus of Social Research Methodology needs to be incorporated into ELSST. The scope of ELSST needs to be widened through the addition of other hierarchies and the use of the partial equivalence relationship has to be investigated further.

Also the CESSDA virtual catalogue needs to develop a thesaurus interface and deal with other multilingual features. Mechanisms, too, have to be implemented for the management of ELSST as a multi-site working tool.

Anyone wishing more information on ELSST should contact Ken Miller at millk@essex.ac.uk.

7. Acknowledgements

We would like to thank all the CESSDA archives who gave us such extensive feedback on ELSST, in particular we would like to thank SIDOS (Switzerland), BDSP (France), ZA (Germany), WISDOM (Austria), FSD (Finland) and CIS (Spain).

8. References

Council of European Social Science Data Archives (CESSDA), http://www.nsd.uib.no/cessda/

Doerr, M. (2001). Semantic problems of Thesaurus Mapping, Journal of Digital Information, volume 1, issue 8.

Humanities and Social Science Electronic Thesaurus (HASSET) thesaurus, http://dasun1.essex.ac.uk/services/zhasset.html

The HEREIN Thesaurus http://www.european-heritage.net/fr/Thesaurus/Contenu.html

ILO thesaurus. (1998). Labour, employment and training terminology, 5th edition, International labour Office, Geneva.

ISO 2788-1986 (1986) Documentation - Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization, Ref. No. ISO 2788-1986

ISO 5964:1985 (1985) Documentation - Guidelines for the establishment and development of multilingual Ref. thesauri, International Organization for Standardization, No. ISO5964-1985

LIMBER project, http://www.limber.rl.ac.uk/

Mérimée thesaurus, http://www.culture.gouv.fr/documentation/thesarch/avertissement.htm

RDF (Resource Description Framework), http://www.w3.org/RDF/

UNESCO thesaurus, http://www.ulcc.ac.uk/unesco/