European Language Social Science Thesaurus (ELSST): issues in designing a multilingual tool for social science researchers

Lorna Balkan
UK Data Archive
University of Essex
Colchester, UK
balka@essex.ac.uk

Taina Jääskeläinen
FSD Finnish Social Science
Data Archive
University of Tampere, Finland
taina.jaaskelainen@uta.fi

Christina Frentzou
EKKE National Centre for
Social Research
Athens, Greece
cfredzu@ekke.gr

Chryssa Kappi
EKKE National Centre for
Social Research
Athens, Greece
ckappi@ekke.gr

Abstract

This paper describes the methodology used to produce the European Language Social Science Thesaurus (ELSST), which has been in development for over decade, supported by a succession of EU-funded projects. Currently available in nine languages, ELSST aims to improve access to comparable social science and humanities data across geography and time. Its design is such, however, that it lends itself both as an information retrieval tool and as a terminological tool more generally.

1 Introduction

Access to good quality data in the social sciences is essential for social and economic policy makers and researchers, and in the European context, this includes in particular access to comparable data across geography and time. The Council of European Social Science Data Archives (CESSDA) operates a data portal which gives access to the data collections of its member states with the aid of a purpose-built multilingual thesaurus. This thesaurus, the European Language Social Science Thesaurus (ELSST), which has been developed over the last ten years and which currently contains nine languages[1], permits users to search for comparable data across different populations using a search term in their own language. There are currently over 3,000 terms for the majority of languages in the thesaurus. This paper explores some of the issues involved in its design and development.

2 Background

Development of ELSST has proceeded under three successive EU-funded projects, namely: Language Independent Metadata Browsing of European Resources (LIMBER), 2000-2003 (Miller and Matthews, 2001); Multilingual Access to the Data Infrastructure of the European Research Areas (MADIERA), 2003-2005; and Council of European Social Sciences (CESSDA)-Preparatory Phase Project (PPP), 2008-2010.

ELSST was initially derived from Humanities and Social Science Electronic Thesaurus (HASSET)[2], the English monolingual thesaurus created by the UK Data Archive, the social science data archive at the University of Essex. Higher level terms from the main HASSET hierarchies were selected in order to arrive at a broader-level, more 'Euroversal' thesaurus, which, it was hoped, would avoid any language or cultural bias. This first phase of ELSST as described in

[1] Lithuanian terms are due to be added to ELSST in spring 2011.

[2] HASSET [5] was originally based on the 1977 UNESCO Thesaurus, ISBN 92-3-101469-2.

Balkan et al. (2002) was confined to English, French, German and Spanish.

In the second phase of ELSST, under MADIERA, four new languages Danish, Finnish, Greek and Norwegian were added[3] and a new methodology introduced. Prior to finding multilingual equivalents to terms, hierarchies were reviewed by a multilingual and multicultural team, and subject experts consulted. Definitions were added to terms where necessary, in order to eliminate further the language and cultural bias inherited from HASSET.

In the latest phase of ELSST, under CESSDA-PPP, a number of hierarchies were amended and enlarged. Earlier translation work had revealed particular difficulties with certain hierarchies, especially education, labour, employment, social welfare and social structure, due mainly to the different systems found in different countries. One solution adopted was to align ELSST terms with international classification systems to deal with these problems.

During CESSDA-PPP maintenance and management procedures were also created, as well as a thesaurus management system.

3 Creating a multilingual thesaurus: the challenges

The first challenge for ELSST lies in the diversity of languages it contains. The second phase of ELSST included the introduction of Finnish and Greek, neither of which belong to the same family as the original ELSST languages (i.e. Romance and Germanic). Finnish in particular is less related to, and has fewer cognates with, the other ELSST languages. While this sometimes makes it more difficult to find Finnish equivalent terms, it avoids the temptation of employing 'false friends', as reported in Jääskeläinen (2006).

A fundamental problem for multilingual thesauri, or for any multilingual language resources, is not only linguistic variation between languages but the fact that different languages have different ways of classifying the world. One language may choose to lexicalise a concept that is lacking in another. Often this is due to cultural differences. For example, Greek has no word for 'house husbands'. Even within the same language (e.g. German), there may be differences in concepts/lexicalisations due to differences in cultural systems such as education and legal systems which may differ between countries and regions. A multilingual thesaurus has to take account of these problems.

Another challenge for ELSST is due to its subject domain, i.e. social sciences. Social science vocabulary has a certain amount of 'hard' terms, i.e. terms which can be precisely defined (e.g. geographical regions), but in the main consists of 'soft' terms, which are much vaguer in scope and which share some overlap with general language. Social science vocabulary thus contrasts with the terminology of the physical sciences, which have a greater proportion of 'hard' terms. Moreover, the meaning of social science terms may vary not just across geographical or cultural boundaries, but across time. An example is 'old age', which means something different today than it did 100 or even 50 years ago[4].

4 Structure and function of a multilingual thesaurus

A thesaurus addresses the problem of vagueness of meaning, in that it is a controlled vocabulary. It consists of a hierarchical arrangement of ('preferred') terms, which express concepts. Terms are intended to express one and only one concept. The relationships between terms are explicitly marked. The hierarchical relationship is the Broader Term (BT) relationship and its inverse Narrower Term (NT). Non-hierarchical relationships include the Used For (UF) relationship, typically synonyms or near synonyms, or antonyms, lexical variants, etc; and the Related Term (RT) relationship, which expresses a looser association to the main 'preferred' term than the BT relationship.

Thesaurus relationships serve several purposes. First, together with the terms they link, they provide a roadmap to the conceptual space of the domain. This can be useful to information seekers who wish to get an overview of the domain or subdomain(s). Second, relationships such as BTs, NTs and RTs can suggest alternative search terms for those using the thesaurus as an information retrieval tool, allowing them to widen or narrow their search. Third, while the relationships between terms in a thesaurus are made explicit, the meanings of the individual terms are frequently only implied, either from their UFs, or from their place in the thesaurus.

[3] Swedish was also added to ELSST at this stage, though not under EU funding.

[4] This is attributable to the nature of the adjective 'old' which is comparative, rather than absolute in value.

Thus ‘courts’ in general language may have several meanings, but its position as an NT to ‘administration of justice’ in ELSST narrows its meaning to legal courts. The definition of a term may also be made precise through the use of a Scope Note (SN). Thus ‘bills’ in general language can have at least two meanings - ‘printed or written statements of the money owed for goods or services’, or ‘proposals for legislation which, if adopted by Parliament, become statutes’. In ELSST, only the second meaning is possible, as the term is assigned an explicit scope note to this effect.

The less ambiguous a term, the more precise it is as an information retrieval tool. For example, if researchers use the ELSST term ‘bills’ to search a database, they will know that the list of documents retrieved will be about legal bills and not any other kind. Contrast this with a free text search, where searching for a term equates to searching for a string, not the concept behind the string, and where the search term ‘bills’ will retrieve instances of any use of the word ‘bills’ not all of which will be relevant.

A third type of non-hierarchical relationship, the equivalence relationship, is found only in multilingual thesauri. This is the relationship which links a term to its foreign language equivalent(s) in the thesaurus. Note that in ELSST, equivalence relationships are always defined relative to the English source term. Given the different ways in which different languages lexicalise concepts, the equivalence relationship may be quite complex, ranging from complete equivalence (where two terms express exactly the same concept) to non-equivalence (where there is no equivalent concept at all in one of the two languages). Five different levels of multilingual equivalence are defined in ELSST, based on Guidelines for Multilingual Thesauri of the International Federation of Library Associations and Institutions:

1. Exact equivalence: source language (SL) and target language (TL) terms refer to the same concept.

2. Inexact or near equivalence: SL and TL terms are generally regarded as expressing the same general concept but the meanings of the terms in SL and TL are not exactly identical. Often the differences are more cultural than semantic, i.e. there is a difference in connotation or appreciation.

3. Partial equivalence: SL and TL terms are generally regarded as referring to the same concept, but one of the terms strictly denotes a slightly broader or narrower concept.

4. One-to-many equivalence: to express the meaning of the preferred term in the SL, two or more preferred terms are needed in the other language.

5. Non-equivalence: No existing term with an equivalent meaning is available in the TL for a concept in the SL, for cultural or linguistic reasons.

It should be noted that ELSST does not aspire to represent all social science concepts, merely those relevant to the existing data collections of the participating archives. Similarly, no formal logic underpins the relations between these concepts - relations such as subtype-supertype or part-whole determine the positions of the concepts in a hierarchy but do not completely define them. Thus, to use Sowa’s (Sowa 1999) terminology, ELSST can be described as a ‘terminological ontology’ rather than a formal ontology.

5 Bridging lexical and conceptual differences across languages

A central problem for multilingual thesauri construction is how to deal with these different types of equivalence relationships between concepts.

Inexact or near equivalence is treated as exact equivalence in ELSST. This is no different in essence to the relationship between a preferred term and its synonyms or near synonyms in the monolingual thesaurus.

Partial equivalence has received different treatments in ELSST. In some cases a BT or NT can be chosen instead. For example, the English term ‘paramedical personnel’ which means persons who work in ambulances and who are trained in first aid, emergency care etc, is mapped to the Finnish term ‘ensihoitohenkilöstö’. The Finnish term is broader in scope, covering, in addition to persons working in ambulances, also those working in emergency care units. In other cases where the meaning diver-

gences are due to culture-specific reasons, and where international classification schemes exist, efforts have been made to import them into ELSST. This is particularly the case for terminologies referring to systems, such as the education, legal or health care system. For example, the International Standard Classification of Education 1997 (ISCED97) was consulted for terms for educational systems and levels. While they offered useful generic terms to describe concepts (e.g. lower secondary schools) they do not necessarily correspond to terms that information seekers would use to search for documents. They thus need to be augmented with country or region-specific UFs (e.g. 'yläkoulut' in Finnish[5] and 'collèges' in French).

An example of single-to-multiple equivalence is the translation of the term 'housewives' into Finnish. The concept of housewives can only be represented by two different concepts in Finnish: 1) 'Kotiäidit' (literally translated 'stay-at-home-mothers' and 2) Kotirouvat (literally 'stay-at-home-ladies'). There is no neutral equivalent of housewives. The two Finnish terms have their own connotations: the first refers to wives staying at home to take care of children (implied by 'mothers') and the second, now becoming old-fashioned, that the family is well-off (implied by 'ladies'). Working class families would not normally have a 'kotirouva'. In ELSST, the equivalence was handled by creating a synthetic term, KOTIÄIDIT JA KOTIROUVAT, which consists of Kotiäidit and Kotirouvat conjoined by 'JA' ('and').

For cases of non-equivalence between languages, several strategies are possible including:

(1) disallow a concept if it does not exist in one or more of the thesaurus languages;
(2) allow the definition of a concept to exist in the thesaurus, without lexicalising it;
(3) adopt a loan word or some other artificial construct as its equivalent.

Strategy (1) is overly restrictive and not an option in ELSST. Similarly (2) is excluded since the structure of each language hierarchy (excluding the number of UFs, which can vary according to language) is identical in ELSST, and every preferred term has to have an equivalent in each of the other languages. Strategy (3) is adopted in ELSST. For example, the concept of 'travelling people' has no equivalence in Finnish, so is mapped to the English term 'travelling people'. From the information retrieval point of view, this is adequate, because a searcher will not be able to find Finnish data about 'travelling people' anyway, since the concept does not exist in Finland.

A novel approach to equivalence problems in ELSST is to adopt a special kind of scope note called the Translation scope note. Thus the case of the difference between 'paramedical personnel' in English and 'ensihoitohenkilöstö in Finnish is explained with the translation scope note both in English: 'The Finnish term covers all personnel with emergency care training working in ambulances or emergency care units', and in Finnish: ' Englantilainen termi kattaa vain ambulansseissa työskentelevät' (the Finnish SN says 'The English term covers only those working in ambulances').

6 Conclusion

Some of the challenges encountered in constructing ELSST stem from the fact that it was derived from an existing monolingual thesaurus, rather than being constructed from scratch (a preferable, but costlier option). The biggest problem is the lack of definitions associated with source terms. It has been necessary to add many more scope notes to the English source terms in ELSST before equivalence relationships could be established.

Another problem is that although discussing and amending English terms and hierarchies in a multilingual and multicultural terms in advance of seeking their multilingual equivalents helps to reduce language and cultural bias, this is not enough for hierarchies describing systems. In this case, there is no alternative to starting from scratch, preferably using international standard classifications and existing thesauri.

Ultimately, it is impossible to eliminate all concept mismatches due to the inherent differences in the way that different languages lexicalise concepts. However, for the information seeker, partial equivalence will in most cases still retrieve relevant data, which is the main purpose of a thesaurus. It is hoped that by adding scope notes, including translation scope notes, these different levels of equivalence will be better understood by the users of the thesaurus, thus enhancing the usefulness of ELSST both as an information tool and as a terminological aid.

[5] This work is currently ongoing and these terms are not yet available on the publicly available version of ELSST.

ELSST is currently available for the general public to view at the following web page: http://elsst.esds.ac.uk/login.aspx. It is envisioned that publicly funded bodies such as university libraries will in future be able to obtain a licence for ELSST, which will allow them to use the thesaurus as an indexing and search tool in their local systems. Anyone wishing more information on ELSST should contact Sharon Bolton at sharonb@essex.ac.uk.

Acknowledgments

We would like to thank all CESSDA archive colleagues who have participated in the construction of ELSST.

References

Lorna Balkan, Ken Miller, Birgit Austin, Anne Etheridge, Myriam Garcia Bernabé and Pamela Miller. 2002. 'ELSST: a broad-based Multilingual Thesaurus for the Social Sciences', proceedings of Third International Conference on Language Resources and Evaluation, Las Palmas.

CESSDA-PPP project, http://www.cessda.org/project/

Council of European Social Sciences (CESSDA) http://www.cessda.org/

International Federation of Library Associations and Institutions Working group on Guidelines for Multilingual Thesauri. 2009. *Guidelines for Multilingual Thesauri*, IFLA Professional Reports 115, The Hague, ISBN 978-90-77897-35-5.

Humanities and Social Science Electronic Thesaurus (HASSET) http://www.esds.ac.uk/search/hassetSearch.asp

Taina Jääskeläinen. 2006. 'Meeting the challenge of a multilingual thesaurus'. Presented at the conference Multilingual Thesauri in Social Sciences, Helsinki.

Ken Miller and Brian Matthews. 2001. Having the right connections: the LIMBER project, *Journal of Digital Information*, 1(8).

Multilingual Access to the Data Infrastructure of the European Research Areas (Madiera) project, http://www.dataarchive.ac.uk/about/projects/past?id=1633

John F. Sowa. 1999. *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing, Co., Pacific Grove, CA.

Standard Classification of Education 1997 (ISCED97),http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm