Title: From monolingual to multilingual thesaurus: the HASSET/ELSST relationship reviewed

Authors: Lorna Balkan

Submitter: Lorna Balkan, balka@essex.ac.uk

Introduction

The UK Data Archive recently received funding in the ESRC-funded CESSDA ELSST project to develop the two social science thesauri which it manages, namely the multilingual thesaurus ELSST and the monolingual thesaurus HASSET from which it is derived. Historically, they have been developed at different rates and on different platforms, which has allowed them to diverge over time. One of the aims of the project was to see whether they could be brought back into alignment, in a new, unified thesaurus management system. The initial hypothesis was that the two thesauri could be merged, i.e. that ELSST could be treated as a subset of HASSET. However, initial alignment work led to the conclusion that it was more appropriate to see the relationship between the two thesauri in terms of mapping, rather than merging. Divergence should be allowed, but strictly controlled. In this paper, we discuss the alignment work we undertook and the mapping relationship that we defined between the two thesauri as a consequence of it. We explore to what extent the differences between the two thesauri can be captured by the axioms and constraints of the new thesaurus management system, and what remains a matter of manual control.

Overview of HASSET and ELSST

HASSET is used as an indexing tool within the UK Data Archive; its primary purpose is to enable users to search the Archive's data collection. ELSST, on the other hand, is primarily used within the CESSDA portal. Both aim to promote cross-national information retrieval in the social sciences. They do this by, amongst other things, providing clear definitions for their concepts, and in the case of ELSST, for their foreign language equivalents.

The main difference between the two thesauri is that, while HASSET may include UK-specific concepts, ELSST must only contain concepts that are common to all (or at least most of) its member countries.

Currently, most ELSST concepts are also in HASSET, although it is allowable for concepts to belong to ELSST, not HASSET, and vice versa. Shared concepts are referred to as 'core' concepts; concepts that belong to either ELSST or HASSET only are referred to as 'non-core' concepts. Non-core concepts in HASSET include, but are not confined to, UK-specific concepts. Whether or not a non UK-specific concept is included in ELSST is a matter for its international translators committee to decide.

Defining the mapping relationship between the two thesauri

In the alignment exercise, all relationships in ELSST that were not in HASSET were identified, and resolved where possible. We concluded that it was not possible to maintain total symmetry between the two thesauri.

We formalised the relationship between the two thesauri as follows. Core concepts are required to share the same:

1. Preferred term label
2. BTs

Where they also have the same scope note (and scope note source) they are labelled 'exact equivalents'. Otherwise, they are 'close equivalents'. They may differ in all other metadata, including narrower terms (NTs) and non-preferred terms (UFs).

These definitions are stricter than the corresponding ISO 25964-2 definitions of equivalence, since they demand identity at structural as well as semantic level. This is necessary to keep track of differences between the two thesauri when they are mounted on the same database and are being managed in tandem. The ISO 25964-2 definition of equivalence, by contrast, is entirely semantic.

The ELSST/HASSET 'exact equivalence' in fact corresponds to 'exact simple equivalence' in ISO 25964-2; we expect any difference between the scope notes of 'close equivalents' in ELSST and HASSET to be small enough for 'inexact simple equivalence' to hold in ISO terms.

Automating the mapping relationship

Currently, the two thesauri function as two separate products, although they have been kept in alignment as far as is possible via a manual process. In the new management system, they will continue to function as two separate thesauri, but consistent alignment will be ensured by mounting them on the same database and imposing a number of constraints on their relationship. This work is underway.

While it is easy enough to ensure that core concepts shall have the same preferred label and BTs, it is more challenging to define how and when core concepts may differ in all other metadata. We anticipate, for example, that scope notes and sources will only differ in core concepts that have some degree of cultural-specificity and/or where the ELSST concept is broader than the HASSET concept.

We further anticipate that, as in the current system, we shall continue to allow a preferred term in HASSET to be a UF in ELSST. For example CHILD BENEFITS is an NT of FAMILY BENEFITS in HASSET but a UF of FAMILY BENEFITS in ELSST. However, again, certain constraints need to be imposed on this relationship.

Conclusion

Work has been carried out to date on aligning HASSET and ELSST in preparation for migrating them to the new management system. The new system should serve the dual purpose of keeping them in alignment, where required, while at the same time allowing them to diverge. It seems likely that the ways in which they may diverge are less easy to formalise, and hence automate, but should reveal interesting facts about the nature of the relationship between the multilingual and monolingual thesauri.