Turning *Roget's Thesaurus* into a Czech Thesaurus

Aleš Klégr
Charles University in Prague

Turning Roget's Thesaurus *into a Czech Thesaurus in a report on how a thesaurus of the Czech language was compiled on the basis of* Roget's Thesaurus*, the following issues are covered:*

1. *Reasons for undertaking the thesaurus project-to redress the unbalance between the semasiological and onomasiological description of Czech by compiling a counterpart to the two large alphabetical dictionaries of Czech;*
2. *Strategy and philosophy, and the choice of the source text-combination of translation and original compilation; decision to use an available and well-proven model, a shorter version of* Roget's Thesaurus*, to resolve the issue of a classificatory system and format;*
3. *Phase one: a project grant-awarded by Charles University for a three-year project, Computerized Thesaurus of the Czech Language, resulting in a preliminary translated version of the Czech thesaurus and the publication of a sample volume as an output;*
4. *Phase two: expanded version for publication-moving from translation to original compilation for greater autonomy of the Czech thesaurus and expanding the average of 80 items per entry to 300 using Czech sources; specific rules required for entry structure, the type and order of subentries, etc, to ensure the uniform format of the entries;*
5. *Compiling the index-to achieve the standard index-length equal to that of the dictionary text, a procedure combining manual and mechanical shortening was devised to abridge the dictionary text;*
6. *Conclusion. Compilation of a thesaurus via translation from another language is a possible procedure. Supplementing translation with original compilation based on target-language resources is nevertheless recommended if a truly national thesaurus is to result.*

The following is a brief report on how a thesaurus of the Czech language was compiled on the basis of *Roget's Thesaurus* and the experience gained therefrom.

Reasons for undertaking the thesaurus project

There are two major dictionaries of Czech, both of which came into being during the 20th century: *Příruční slovník jazyka českého* (1935-1957) and *Slovníku spisovného jazyka českého* (1960-71). The former has some 250 000 entries, the latter one fifth fewer, i.e. 192,000 entries. There is no onomasiological dictionary for Czech corresponding in size or scope of reference. Currently there is one small-sized dictionary (Pala & Všianský 1994) of synonyms featuring some 40,000 synonyms, which is only a modest improvement on its two predecessors (Mašín & Bečka 1947; J. V. Bečka 1982). There is only one thematic dictionary of Czech, based on Hallig and Wartburg's classificatory system, Haller et al., *Český slovník věcný a synonymický,* (1969-1986). Unfortunately this ambitious project has remained uncompleted, covering only three topics (nature and the physical and psychological aspects of man). The aim of the thesaurus project was thus to attempt to compile a counterpart to the two large alphabetical dictionaries of Czech and redress the unbalance between the semasiological and onomasiological description of the Czech language. When looking for a suitable model, the choice was predetermined partly by the author's specialization in English linguistics, the outstanding record of *Roget's Thesaurus* (celebrated by the 150th anniversary edition) and the success of the thesaurus format in other languages, such as German (Dornseiff 1954), Spanish (Casares 1959), Dutch (Brouwers 1965) or French (Péchoin 1995).

Strategy and philosophy, and the choice of the source text

The approach had to be pragmatic because of technical, financial and staffing limitations and, given the conditions, fairly traditional. The idea of "[A]utomatic thesauruses", produced by processing corpora, with similarity between words measured (directly or indirectly) by co-occurrence', despite Kilgariff and Yalop's apparent optimism (2000), was not a feasible solution.

When contemplating the task of thesaurus compilation, two of its aspects have to be considered: the system of classification on which it is to be built, and, second, the strategy of writing the dictionary articles (entries). Experience shows that trying to develop an original plan is a lengthy and laborious business with uncertain results. Conversely, if there is already such a plan and clearly functional at that, it makes sense not to look much further.

As regards the strategy of writing entries, once the classification has been decided upon, there are two options. The first is to use not only the classification, but also the actual articles of the source thesaurus and translate them into the target language. The second is to use the adopted classificatory system of (sub)heads and use them only as the starting point, while disregarding the text of the entries. Both have its pros and cons.

Translation is seemingly the easier solution. However, with the enormous of number of lexical items out of context, from different stylistic levels and often language-bound and culture-specific, the advantage of having ready-made entries may easily dissolve in the process of struggling with tricky translation problems. Interestingly, many people found the idea of "translating a thesaurus" unrealistic, though they could not easily explain why. Presumably, they considered this kind of dictionary to be far too language-related, although Peter Mark Roget himself anticipated a bilingual thesaurus and thought such a "conjunction" useful (cf. Lamy, Towell).

The second approach, fleshing out the classificatory system from scratch, may involve difficulties of a different type. You need to figure out the logic and niceties of the semantic structure of each entry which may not be immediately obvious from the list of the (sub)heads. Also, the role of the compiler changes substantially, as he suddenly becomes more of an author than a translator. Apart from requiring different qualities in the compiler, such a switch entails different strategies in selecting and arranging the items to be included and a different type of literature resources.

There is, however, a third possibility, the combination of translation and original compilation. The translation phase provides the blueprint and indicates directions in which the next stage of expanding the article text may go. As it happened, the implementation of the project went through two phases which made it possible to apply both strategies, translation and compilation with little or no influence from English.

As mentioned above, the choice of *Roget's Thesaurus* as the source was influenced by its continuing success, universal structure and variety of semantic relations it covers, and partly by the affinities with Comenius' *Janua Linguarum Reserata*. The decision to use an available and well-proven model resolved the issue of a classificatory system (word list and its structure of opposing concepts), the format of entries (subdivided according to word-class) and that of the index, and made it possible to concentrate exclusively on the implementation.

The next step was to select from among the many editions and modifications of *Roget's Thesaurus*. Of the three distinct lines of thesauri, British (Longman, Penguin, etc.), American (Crowell, Harper-Collins) and Australian (Macquarie), the British one was chosen. After considering the classical British thesauri on offer, we opted—perhaps surprisingly—for *The Penguin Pocket English Thesaurus* (Carney, Waite, 1985). Compared to the full versions, this one uses fewer categories/heads (882 instead of the standard 990, or the original 1000) and contains fewer lexical items (70-80 000, compared to some 320 000 items in the full-sized versions).

Again, the choice had its merits and demerits. On the plus side, the system and order of sections/heads, although reduced (or rather telescoped), preserves the original plan of *Roget's* (as modernized by Robert A. Dutch) intact. The important advantage of this "skeleton" edition is that being stripped to the basics it allows the compiler greater leeway for transposition to Czech and presents fewer translation problems. Also, the prospect of having to deal with "only" 80 000 lexical items rather than 320 000 is much easier to face.

On the minus side, the reduction (or coalescence) of many sections/heads in this smaller thesaurus may to some extent affect the internal structure of the entries which then becomes sometimes less focused and clear-cut than in unabridged thesauri.

Phase one: a project grant

The starting-point was a project grant awarded by Charles University, Prague, for a three-year (1999-2001) project called Computerized Thesaurus of the Czech Language. Its goal was to create a text on the basis of the English thesaurus for a Czech thesaurus database convertible into a printed version. It represented the translation phase of work on the thesaurus and served as a feasibility test of translating a thesaurus from one language into another. It allowed engaging part-time collaborates, in fact undergraduate students of English, who provided about two thirds of the draft translation. Although MA-level students of English, they generally found the translation rather difficult and accordingly the standard of the translation varied in quality. The difficulties were identified as due to translation of words out of context, the extensive polysemy in English and the failure to maintain the semantic tenor of the entries.

By 2001 the complete translated version and the revision of text and format of some 40 % of the entries had been finished. The editing included collation with the word list of the standard medium-sized school dictionary of Czech (*Slovníku spisovné češtiny pro školu a veřejnost*, 50 070 lexemes) with 223 selected "cleaned-up" entries of the sample volume. These entries consisted of 30 760 lexemes (44 000 tokens) and comparison with the Czech alphabetical dictionary word list showed concurrence in 13 800 lexemes. In other words, the sample volume of the Czech thesaurus exhibited a 28 % overlap with the word stock of the 50 000-lexeme Czech dictionary. This was taken as a good sign of the lexical representativeness of the thesaurus, considering the entries were still in the translation stage. Alongside the translation, an on-line electronic database was prepared, necessary for the production of the sample volume and the generation of its index.

The project, concluded by the publication of the sample volume, demonstrated that it is in principle possible to translate a thesaurus from one language into another. On the other hand, it also showed a certain amount of language-specific and cultural asymmetries and gaps that had to be compensated for and accordingly a certain degree of language interference. A case in point is the translation of the 882 heads of the English source thesaurus. The translation of 35 of them resulted in duplicities in Czech (the same word appeared twice), which was partly due to lack of coordination between translators, partly to lexical reasons. Also, some difficulties in translation were encountered when the editors of the English thesaurus let themselves be influenced by the associations of English words going beyond the conceptual classification.

Phase two: expanded version for publication

Although phase one went generally quite well, it also showed there is scope for further improvement. If anything the text was still to a large extent a draft translation. Phase two aimed at greater autonomy of the Czech thesaurus from the original through expanding the average number of 80 items per entry to 300 using Czech sources.

This decision had some inevitable consequences. The shift from translation to original compilation precluded the participation of inexperienced undergraduate students. In fact the work proved to be so specific and time-consuming that in the absence of any further grant funding, the team shrank to a single author. The increased size of the entries increased the amount of labour and time required to finish the manuscript (stretching to six more years), and required that certain rules for entry structure, the type and order of subentries and their content be specified and observed in order to ensure the uniform format of the entries. In nominal groups (subentries) the typical order is abstract noun (quality, property), abstract noun (activity, act), concrete noun (object), concrete noun (agent, bearer), and concrete noun (patient). Adjectival and verbal groups follow suit as far as possible. The ordering of items within subgroups is not alphabetized, but rather tries to follow the inner logic of the concept unfolded (vernacular words followed by the foreign ones, etc.).

Although the subdivision of the entry into nominal (sb.), adjectival (adj.), verbal (vb.) and adverbial (adv.) sections is very useful, there are lexical items that deserve inclusion but do not fit any of these parts (such as comparisons, proverbs, quotations, etc.). In such cases the adv. section was used as a catch-all. The order of heads (entries) remained unchanged. Only in three cases was it thought expedient to add heads: 439 Science, 763 Sport and 865 Gods, Deities (by dividing the head 864 Divinity into two, as in the unabridged version). The number thus increased to 885 heads.

The translated text of the phase-one thesaurus with the subentry heads as pointers provided the guidelines for the compilation based only (or mainly) on Czech sources. The sources included the electronic versions of the largest contemporary dictionary of Czech available (*Slovník spisovného jazyka českého*), and its medium-sized derivative (*Slovník spisovné češtiny pro školu a veřejnost*), a dictionary of foreign words (*Velký slovník cizích slov*), three Czech electronic encyclopaedias (*Encyklopedie Diderot, Encyklopedie Universum, Encyklopedie Heuréka*) and a number of printed reference books (such as dictionaries of Czech neologisms, colloquial and slang expressions, terminological dictionaries, etc.). An important source of information was both the abridged and unabridged corpus of the Czech language (*Synek/Litera* and *Syn2000*).

Compiling the index

The final but crucial step was the compilation of the index. When the typical index format is observed (4 vertical columns, smaller font size, target word on a separate line, word-class information unnecessary in Czech), the transformation of the full dictionary text into an index via the database inevitably results in a text three to four times longer. As the customary length of a thesaurus index is roughly the same as that of the dictionary, some kind of abridgement is clearly called for. Unfortunately the procedure and principles for such a reduction are not readily available and so they had to be devised for the occasion.

The first important decision is whether to make cuts first in the text and then compile the index or first compile the index and then start abridging it. Whatever is chosen, the cuts need to take place in the database. The second decision is whether the cuts should be made manually or mechanically. And, finally, it is important to know the extent of shortening necessary (by a third, a half?) to produce an index of the right size. As time was a key factor, the procedure had to be as time- and labour-saving as possible and so we opted for a combination of manual and mechanical abridgement. The manual shortening was made in selected entries; the mechanical shortening involved three steps, each of them carefully designed (a) to preserve items that should stay if the index should remain functional, and (b) to allow for controlled reduction in size. The finishing phase is that of layout and font-size manipulation to produce an index of the required length. The outcome was surprisingly good given the time available and lack of experience.

Conclusion

The fact that the project reached the stage of being accepted and published by a commercial publisher suggests that compiling a thesaurus via translation from another language is a possible procedure. However, supplementing translation with original compilation based on target-language resources is recommended if a truly national thesaurus is to result. Naturally, this leads to a departure from the original and accordingly requires a greater lexicographic input on the part of the compilers.

References

Bečka, J. V. (1982). *Slovník synonym a frazeologismů*. 3. vyd. Praha: Novinář.

Brouwers, L. (1965). *Het Juiste Woord*. Antwerpen-Utrecht: Standaard Uitgeverij.

Casares, J. (1959). *Diccionario ideológico de la lengua española*. Barcelona: Gustavo Gili.

Dornseiff, F. (1954). *Der Deutsche Wortschatz nach Sachgruppen*. Berlin.

Haller, J. a kol. (1969, 1974, 1977, 1986). *Český slovník věcný a synonymický 1, 2, 3*. Praha: Rejstřík, SPN.

Hallig, R.; Wartburg, W. von (1963). *Begriffssytem als Grundlage für die Lexicographie. Versuch eines Ordnungsschemas*. Berlin.

Hüllen, W. (2004). *A History of Roget's Thesaurus. Origin, Developments and Design*. Oxford: Oxford University Press.

Kilgarriff, A.; Yallop, C. (2000). "What's in a thesaurus?". In *Proceedings of Second Conference on Language Resources and Evaluation*. Supported by Macquarie University Visiting Researcher Fund. 1371-1379.

Lamy, M.-N.; Towell, R. (1997). *The Cambridge French-English Thesaurus*. Cambridge: Cambridge University Press.

Macquarie Thesaurus, The (1984, rev. 1986, reprint 1992, red. J. R. L. Bernard; rev. 2000, red. J. R. L. Bernard, S. Butler). Macquarie Library Pty Ltd. & Macquarie University NSW.

Martincová, O. a kol. (1998, 2004). *Nová slova v češtině. Slovník neologizmů*. Praha: Acaemia.

Mašín, J.; Bečka, J. V. (1947). *Malý slovník českých synonym, nakl*. Praha: Mikuta.

Pala, K.; Všianský, J. (1994). *Slovník českých synonym*. Praha: NLN.

Péchoin, D. (1995). *Thésaurus*. Paris: Larousse.

Roget's International Thesaurus (1886). Thomas Crowell Company [5th 1992 (ed. R. L. Chapman, Harper Collins).].

Roget's Thesaurus of English words and phrases. London: Longman, 1962. Ed. Robert A. Dutch [2002 (ed. G. Davidson, Penguin).]

Slovník spisovné češtiny pro školu a veřejnost. 2. vyd. Praha: Academia, 1994.

Slovník spisovného jazyka českého. 2. vyd. Praha: Academia, 1989.

The Penguin Pocket English Thesaurus. London: Penguin Books, 1985.

Velký slovník cizích slov. Elektron. vyd. Praha: Academia-Leda, 1994.