Klaus Schubert
Fachhochschule Flensburg, Germany

Parameters for the Design of an Intermediate Language for Multilingual Thesauri

Klaus Schubert was a researcher in sociolinguistics at the University of Kiel, a computational linguist in the software house Buro voor Systeemontwikkeling at BSO/Research in Utrecht, a consultant at BSO/Language Technology in Baarn (Netherlands) and is at present professor of computational linguistics and technical translation at Fachhochschule Flensburg.

Schubert, K.: **Parameters for the design of an intermediate language for multilingual thesauri.**
Knowl.Org. 22(1995)No.3/4, p. 136-140, 9 refs
The architecture of multilingual software systems is sometimes centred around an intermediate language. The question is analyzed to what extent this approach can be useful for multilingual thesauri, in particular regarding the functionality the thesaurus is designed to fulfil. Both the runtime use, and the construction and maintenance of the system is taken into consideration. Using the perspective of general language technology enables to draw on experience from a broader range of fields beyond thesaurus design itself as well as to consider the possibility of using a thesaurus as a knowledge module in various systems which process natural language. Therefore the features which thesauri and other natural-language processing systems have in common are emphasized, especially at the level of systems design and their core functionality. (Author)

1. Thesaurus Design as Software Engineering

The decision whether or not to use an intermediate language in a multilingual thesaurus has to be made during systems design. Since in our days thesauri could be characterized as a sort of software systems, the design in question is one of the phases in software engineering - with the special complexity that characterizes linguistic software design.

In software engineering it is a common procedure to let a systems analyst, in cooperation with the customer or the user, work out a *functional design* of the system to be developed, which is then further processed by a systems designer to yield a *technical design*. This is in turn broken down into modules which are implemented by programmers. Due to the extraordinary complexity of advanced systems in language technology, it is often advisable to lay a third level above these two, the level of *linguistic design*.

This article offers some general considerations on the design of multilingual thesauri. In order to present these without getting into the technicalities of particular systems, the issue of an intermediate language mainly at the level of linguistic design is discussed.

2. Elements of Thesaurus Design

A thesaurus resembles encyclopaedias, dictionaries and term banks in that it describes words. Its distinguishing characteristic is the order in which it arranges the words. The ordering criteria include conceptual or semantic relations as well as relations of an extralinguistic nature. In the definitions of systems theory a thesaurus (whether or not realized as software) is a system, since it consists of elements and the relations between them. Thesaurus design has to take into account both the elements and the relations and define these with respect to the envisaged application of the thesaurus.

This leads to the question of the purpose which the thesaurus is designed to serve. I shall not deal here with the benefits and application fields of thesauri, but restrict my discussion to the two major areas of application which need to be distinguished from a point of view of linguistic design:

Thesauri are used by people for reference. A core function is the standardization of conceptual relations and categorizations which are known to be neither intuitively obvious nor inter-subjectively unambiguous, so that they have to be prescriptively defined.

Thesauri are used by natural-language processing systems as a knowledge source. Thesaurus knowledge is needed in machine translation, in meaning-based information retrieval, in automatic summarizing of texts, in relevance ranking, in information routing and related fields. I do not mention artificial intelligence in this list; it plays its role at a different level, furnishing the instruments that make thesaurus knowledge applicable to the tasks encountered in those systems.

These considerations apply to thesauri in general. Thesauri need to be *multilingual* wherever they have to support work in several languages in parallel, as for instance in industrial documentation, manual or machine translation, international standardization or related areas.

3. A Language-independent System of Concepts?

The idea of addressing the design of a multilingual thesaurus with the tools of general language technology may well be appealing. It is a prerequisite for such an approach, however, that one should be able to word in the terms of language technology what thesaurus specialists, albeit in other wordings, have defined as the functional requirements to thesauri.

When I say in section 2 that a thesaurus describes *words*, this must offend the ears of thesaurus specialists and terminologists. They normally stipulate that their

definitions describe *concepts* which are then labelled with designations. This way of viewing things includes the assumption that concepts are defined cross-linguistically, i.e. in a way equally valid for several languages, and that the labellings in the individual languages are then hooked up to this language-independent system of concepts.

For the decisions to be taken when designing a multilingual thesaurus, the question whether such a language-independent system of concepts or a cross-linguistic semantics is possible in theory and feasible in practice, is of central significance. Before the benefits and drawbacks of an intermediate language can be discussed, this question needs some further attention.

There are three objects to the idea of a language-independent system of concepts as far as the discussion is here concerned, two of them of a practical, language-technological nature and one of them theoretical:

a) *Practical objection*: A language-independent concept cannot be written down. It is only possible to write down words taken from a language. Thus a thesaurus contains words from some sort of language.

It is possible and indeed usual, one could counter-argue, to differenciate the words in thesauri further than the language does ($ball_1$, $ball_2$, $ball_3$). Moreover, it is feasible to define new concepts and represent them by means of some string of characters, which, it is normally claimed, are words from an existing language only as a simple measure of comfort (**BALL-TOY**, **BALL-DANCING PARTY**). However, it is not sufficient to define concepts as isolated symbols. What is needed is a whole, coherent *system* of symbols, i.e. concepts and the relations that hold among them. One of the great authorities on language theory, Louis Hjelmslev (1963: 101), tells us (in my words) that an artificial symbol system is inherently less expressive than a human language:

b) *Theoretical objection*: A *system* of artificially defined concepts is a subset of the concept system of a human language. Its expressiveness is lower than that of a human language. A symbol system with the same degree of expressiveness as a human language is a new language. It cannot be artificially devised.

There is another objection from language technology, which is not concerned with the possibility or impossibility of a language-independent concept system, but with a practical obstacle:

c) *Second practical objection*: In language technology it is highly desirable to enable *automatic* knowledge acquisition. In the case of thesauri, this means acquiring the elements and relations that make up a thesaurus structure from corpora, which are texts in human languages. There is no such thing as a huge amount of undoctored texts written in a symbol system which the systems designer has just artificially devised.

The idea of language-independent concepts seems to encounter so many difficulties, not least where the practical aspects of its realization in language technology are concerned, that it is worthwhile re-thinking the question whether this really is the only possible way. There might be other ways of achieving the same effect.

In my view a good point of departure can be found in the assumption that it would make sense as a first step to develop a general system of concepts and relations between them, which would be valid for all languages. The individual words of the various language would then need to be hooked up to this system in a straightforward manner. This assumption implies the requirement of straightforward mapping of the semantic systems of various languages onto the concept system to be developed and thus, transitively, onto each other. Machine translation, however, has learned from several decades of painful experience that the semantic systems of human language can by no means be mapped onto each other straightforwardly. Despite all successes in other areas, machine translation systems still stand before the same semantic barrier which had been encountered in the first research attempts as soon as they reached a somewhat realistic size (cf. e.g. Yngve 1967: 500).

To claim that there should be a relationship of straightforward mapping between a concept system to be created and existing human languages, means therefore in essence, that the concept systems of these languages are adapted to the human-made concept system. Indeed, active interference with a linguistic system is the proclaimed goal of terminological standardization. The experience of standardization (as well as that of language planning) shows, however, that it is not possible to arbitrarily interfere with the semantic system of human languages. Languages make up independent semantic systems which do exercise some influence on each other from time to time, but which most of the time develop autonomously. Linguistic systems are to some very limited extent susceptible to the active interference of terminologists and language planners, but mainly they develop according to their own laws in an essentially anarchic way.

It may be objected that thesauri are normally developed for specialized language which is more open to active interference than common language. Yet, this is only a gradual, not an absolute difference. The basic observation that there is no straightforward mapping between the concept systems of any two languages, holds for specialized language as well, albeit to a somewhat lesser degree. The fact that this impossibility cannot by any means be removed is due to a number of factors:

– Specialized and common language cannot be neatly separated. The core vocabulary of all texts, whether or not in specialized language, belong to the basic stock of common-language words. In addition, frequently used terms tend to float from specialized into common language.

– When thesauri are used in natural-language processing applications, it is desirable to cover the entire language in a single type of knowledge source, rather than building up, for instance, a thesaurus which cannot then be applied to the common-language words.

– Thesauri are often concentrated on nouns, which make up the overwhelming part of the specialized vocabulary. However, the relations, which by definition are the constitutive characteristic of a thesaurus, are found mainly in verbs and adjectives (more precisely: in predicating words). Automatized knowledge acquisition, that is, the automatic acquisition of *both* concepts *and* relations, should therefore have access to the entire vocabulary, and in particular to the verbs of the core vocabulary, thus to those elements that are not part of the specialized language.

The idea that relations should be recognized and acquired automatically establishes a special link to grammar models that describe the syntactic relations directly, thus dependency grammar in particular. This link is recognized in various works on thesaurus design.

Natural-language processing applications are generally limited to working with the form side of the linguistic sign in order to simulate any processing of the content side. In much the same way it is necessary to acknowledge that in thesauri one is limited to working with words in order to represent a processing of concepts. This insight is fundamental to the application of an intermediate language.

4. The Role of the Intermediate Language

The role that the intermediate language can play in a multilingual thesaurus is closely linked to the above considerations concerning a language-independent system of concepts. Two essential questions arise: Before opting for an intermediate language, one should realize what the special benefits are which are offered by the intermediate language and which could not be obtained without it. If that consideration leads to a decision in favour of an intermediate language, one should ask which properties the intermediate languages should have. Linguistic systems design includes these two considerations.

First, let us consider the question of the function of an intermediate language in a multilingual thesaurus. In my view, its basic function is to express the *definitions* of concepts so that they need not be repeated in more than one language. The basic function also includes the *classification*, that is, the entire set of relations is defined in the intermediate language without resorting to any other languages. Transferring the definitions of concepts and relations into the intermediate language leads to a most interesting corollary: the automatic acquisition of concepts, relations and possibly even definitions should by preference be carried out in the intermediate language as well.

Which are the special advantages offered by an intermediate language in view of these functions? First of all, there is a specific reasoning that underlies the description of the function of an intermediate language as given above: If it is true that concepts cannot be defined in a language-independent way, the designer may at least opt to define them in a single, carefully chosen, language which will then play a pivotal role in the thesaurus. The intermediate language becomes the sole medium of concept definition. This does not yield language-independent definitions but it offers the advantage of a uniform definition valid for all languages used in the thesaurus. However, the possibly complex problem of mapping the semantic system of the intermediate language onto those of the various thesaurus languages is not removed. Defining concepts exclusively in the intermediate language further entails the advantage of decreasing the amount of work required for establishing the required definitions and it contributes to higher modularity and consistency within the thesaurus. As thesauri are often enormously large, these are tangible advantages. A series of additional factors determine whether these advantages outweigh the effort of introducing an intermediate language.

Before addressing these factors, however, one should consider whether a thesaurus with concept definitions in a single language only can fulfil the functions for which it is devised. For direct user access there should be no major problems as long as one can make sure that those users who have access to the definitions understand the intermediate language. For the use of a thesaurus as a knowledge source for other natural-language processing systems evidence is found in the experience from machine translation, a field where intermediate languages and representations have been an important issue for decades. In this case semantic definitions are not only needed in a well-worded form as found in terminological dictionary books, but at the same time in a form of representation which can be utilized by other natural-language processing modules. This representation may take the shape of semantic features (an instance of explicit definitions) or the shape of contexts (an instance of implicit definitions). Even if definitions are expressed in a formal notation, they are based in a specific human language, as shown in the above argument about the impossibility of language-independent definitions. Even definitions formalized in this way are definitions worded in a specific language, the intermediate language.

As an example, thesaurus information can be used as the data base for a semantic decision mechanism. The question at hand then is whether the semantic decisions needed e.g. for lexical transfer in machine translation from English into French can be made on the basis of thesaurus knowledge that is neither in English nor in French, but exclusively in the intermediate language. Most interesting evidence for this type of approach is found in the *Distributed Language Translation* (DLT) project. DLT was an industrial research and development effort which between 1984 and 1990 developed a proto-

type of a machine translation system with an intermediate language. The first prototype of the DLT system, finished in 1987, contained a semantic module which was fully congruent with the approach described above: translation from English into French through the intermediate language Esperanto by means of semantic information exclusively written in Esperanto. When performing the lexical transfer from the intermediate language into French the system made use of disambiguating semantic information written solely in the intermediate language (Sadler 1989a: 15-106, 1989b: Schubert 1988a).

The DLT prototype gives rise to yet another consideration. In thesauri it is common to name, class and label the semantic relations between the various concepts, which yields a semantic network. A good deal of the relevant research literature is concerned with the choice of the best or the most suitable system of such relation labels. (This is normally worded in more sophisticated terms. In essence, though, it most often boils down to a question of labelling.) Having tried out a series of solutions involving labelled semantic relations, the DLT research project gave up this approach in favour of a different direction of development which prepared and in part realized automatized knowledge acquisition. In textual knowledge acquisition a considerably higher degree of automation can be achieved, when relations are not distinguished by arbitrarily chosen labels, but by linguistic means that can be found directly in corpus texts. Rather than assigning names to semantic relations, the new DLT solution used verbs, prepositions and other function words and function morphemes from the text itself. This solution thus makes use of an *implicit semantics*, whereby it differs slightly from what is usual in thesaurus design. For machine translation, only the functional result counts (along with software engineering criteria such as maintainability, inspectability etc.). In a thesaurus, which normally does not serve only other software modules but also or solely human users, it may become important to render the relation labels in a form the user can read, which need not be a trivial transformation. It stands to reason, however, that solutions geared towards *automatic knowledge acquisition* or at least a high degree of automation in knowledge acquisition strongly suggest a preference for *linguistic means* for the labelling of relations (and concepts). This insight, which is underpinned by experience from various fields of language engineering, may be valuable in thesaurus design as well.

Whether an implicit semantics in the intermediate language is sufficient, depends on the envisaged application of the thesaurus. There is a series of both human activities and machine functions that need a thesaurus as a knowledge source. Whenever the purpose is not only a static parallelism among several languages, but, as in the case of machine translation, a dynamic transition from one language to another, the semantic system will not achieve the required degree of precision and reliability in making knowledge-based context-sensitive decisions only by means of a semantics in the intermediate language. Much better results can be achieved with a semantics that can make use of implicit knowledge of *correspondence relations* between the intermediate language and the source or target language in questions (Sadler 1989a: 110-116). The reason for this can be found in the fact that the systems of semantic relations in different languages are incongruent, so that different languages are more than different sets of words for the same entities, properties and events. Therefore, when one tries to keep semantics implicit, semantic information is needed for a language *pair*. If, however, one chooses to restrict semantic information to the intermediate language, it becomes inevitable to express in the intermediate language even semantic distinctions which the intermediate language itself does not distinguish linguistically. In other words, one would make explicit in the intermediate language the ambiguities of the source and target languages connected to it. This leads to an explicit semantics which in a multilingual system entails the danger of combinatorial explosion or a so-called exploding intermediate language.

The choice thus is between an *explicit semantics in the intermediate language alone* and an *implicit language pair semantics*. In the latter solution - which suggests itself in view of automatic knowledge acquisition - the intermediate language will always be one of the two languages in a pair, thereby linking all languages in the system in a modular way.

It should be borne in mind that an intermediate language in a multilingual thesaurus cannot deliver something which would be incongruent with linguistic facts: it cannot force different systems of semantic relations into a single uniform one. Because of this, the intermediate language should not be taken as a uniform cross-linguistic semantic representation during systems design. The function which the intermediate language in a multilingual system can fulfil is a different one: *the intermediate language can map the semantic systems of different languages onto each other*.

5. The Properties of the Intermediate Language

When specific requirements have led the linguistic systems designer to opt for an intermediate language, the question arises as to what the properties of that language should be. As is in the case of the question of whether or not an intermediate language is needed at all, in this case, too, the answer depends on a number of conditions which may differ for different systems. Rather than answering the question here, I therefore discuss the conditions that suggest specific solutions.

Whether a language or another symbol system is suitable for the function of an intermediate language, can be assessed by means of criteria, which must include the following minimal set. An intermediate language:

- must be able to express all concepts of human thought.
- must be able to express all semantic relations of human thought.
- should facilitate automatic analysis.

– should facilitate automatic knowledge acquisition.

The following systems (ordered on an axis from natural to artificial; cf. Schubert 1989: 22-23) are possible intermediate languages:
– ethnic languages such as English, German or French,
– planned languages such as Esperanto,
– artificial symbol systems such as the semantic representations in artificial intelligence or the intermediate representations in machine translation.

Set off against the criteria mentioned above, the following advantages and disadvantages can be detected in these three kinds of possible intermediate languages:

Ethnic language

Ethnic languages have the full expressiveness required. Automatic analysis (parsing and semantic interpretation of parsed texts) is expensive and cannot be carried out fully automatically. Knowledge acquisition cannot either be fully automatic due to the problems in parsing and disambiguating. Text material in specialized language, the most frequently needed raw material for knowledge acquisition, is normally available. However, when a thesaurus is to be built up for a new field of technology, it is not always certain that a reasonable volume of relevant text material can be found in every ethnic language.

Planned language

Of about one thousand projects of planned languages suggested to date, interlinguistics normally acknowledges only a single one as a real language. Only Esperanto has until now fully passed the transition from an artificial symbol system to a human language, so that it possesses the full expressiveness needed. In the sense of Hjelmslev's hypothesis Esperanto was, when launched, an artificial symbol system and as such dependent on ethnic languages and insufficiently expressive. It is only by unreflected use in a (second) language community for over a century that Esperanto has become a human language in the Hjelmslevian sense of the word. Only for this reason may Esperanto be considered today as an intermediate language (Schubert 1988b, 1992).

Automatic analysis is less expensive in Esperanto than in ethnic languages. Being a language spoken by humans, however, Esperanto has a syntactic structure which is not unambiguous in the strict sense of a parsing algorithm. In knowledge acquisition a considerably higher degree of automation than in ethnic languages can be achieved because of high syntactic clarity and far-reaching semantic compositionality (Schubert 1993). Text material is abundantly available as far as common language is concerned whereas in specialized fields corpus material may be scarce. They can, however, in many cases be written if needed.

Artificial symbol system

An artificial symbol system cannot achieve the full expressiveness of a human language. It will always remain a true subset. Analysis can be fully automatic and perfect if the system is accordingly designed. The same holds for automatic knowledge acquisition which, however, is of little use, since there are no large volumes of text in artificial symbol systems.

This overview may show that the established criteria, though not totally excluding one another, contradict each other. There is no such thing as the one and only, absolutely best intermediate language. Rather, there is a trade-off based on a variety of criteria, among which the one or the other criterion may be preferred by the specific conditions of a given case. This is one of the decisions to be made in linguistic systems design.

6. Diagnosis

In the architecture of a multilingual thesaurus an intermediate language *may* be a useful instrument. It is one of the goals of linguistic systems design to define the precise tasks and functions of the intermediate language as dictated by the specific requirements of the system at hand and to decide whether the advantages of an intermediate language in the given case outweigh the effort of introducing and maintaining it. For a thesaurus of a certain size, the planning phase should include a *feasibility study* to assess the efficiency of various possible intermediate languages on the criteria of the established tasks and functions.

References

(1) Hjelmslev, Louis (1963): *Sproget*. København: BerlingskeForlag
(2) Sadler, Victor (1989a): *Working with analogical semantics*. Dordrecht/Providence: Foris
(3) Sadler, Victor (1989b): Knowledge-driven terminography for machine translation. In: Schubert, K. (Ed.): *Interlinguistics*. Berlin/New York: Mouton de Gruyter, 311-335
(4) Schubert, Klaus (1988a): The architecture of DLT - interlingual or double direct? In: Maxwell, D., Schubert, K., Witkam, T. (Eds.): *New directions in machine translation*. Dordrecht/ Providence: Foris, 131-144
(5) Schubert, Klaus (1988b): Ausdruckskraft und Regelmäßigkeit: Was Esperanto für automatische Übersetzung geeignet macht. *Language Problems and Language Planning* 12: 130-147
(6) Schubert, Klaus (1989): Interlinguistics - its aims, its achievements, and its place in language science. In: Schubert, K. (Ed.): *Interlinguistics*. Berlin/New York: Mouton de Gruyter, 7-44
(7) Schubert, Klaus (1992): Esperanto as an intermediate language for machine translation. In: Newton, J. (Ed.): *Computers in translation*. London/New York: Routledge, 78-95
(8) Schubert, Klaus (1993): Semantic compositionality. *Linguistics* 31: 311-365
(9) Yngve, Victor (1967): MT at M.I.T. 1965. In: Booth, A.D. (Ed.): *Machine translation*. Amsterdam: North-Holland, 451-523

Address: Prof. Klaus Schubert, Studiengang Technikübersetzen, Fachhochschule Flensburg, Am Bundesbahnhof 1, D-24937 Flensburg, Germany.
Tel. +49461 14497-12, Fax +49 461 21125, Internet: schubert@fh-flensburg.d400.de