TOWARDS A CONCEPT OF MULTIMEDIA LINKABLE REFERENCE LEXICONS

CARLA MONTEZ FERNANDES
(Vrije Universiteit Amsterdam)

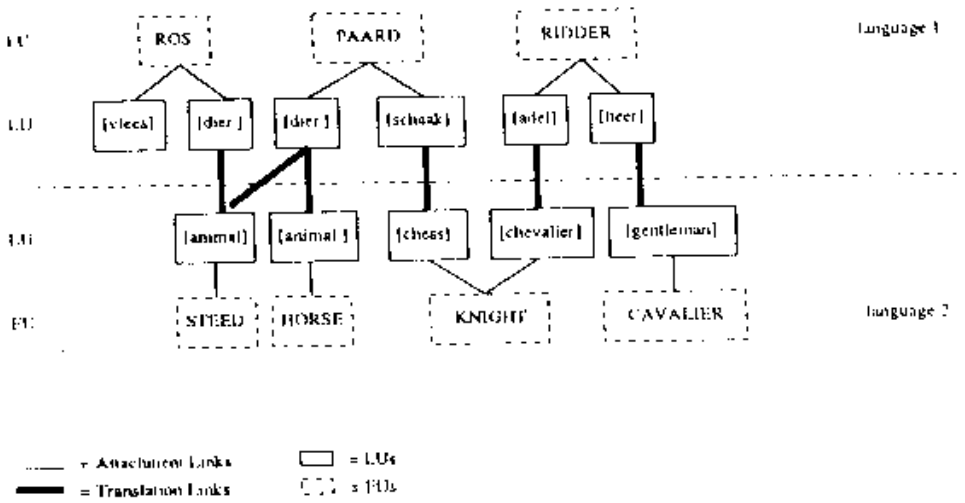
This article is intended as a brief overview of a project for Linkable Reference Lexicons (**LRL**), under development at the *Vrije Universiteit* of Amsterdam, Department of Lexicology, and it will focus on its main execution lines.

The project's starting point is the compilation of a new **Reference Lexicon** for the Portuguese language. This Reference Lexicon should be a multifunctional database, i.e., a non-user oriented lexical database, where several forms and formats of dictionaries can be derived from, as opposed to "front-end dictionary products". The concept of a **Linkable** Reference Lexicon refers to the possibility for this database to be linked to another language (A>B), where its reverse part (B>A) can be (semi-) automatically derived from the original bilingual A>B part.

Ideally, a reference lexicon should contain the following seven levels of information and data categories: graphemics, phonetics, morphology, syntax, semantics, pragmatics, combinatorics, whereas a linkable reference lexicon should provide fine-grained information on syntax, semantics, pragmatics and combinatorics.

LRLs must contain **FUs** (form units) and **LUs** (lexical units or meaning units) and two types of links:

- *attachment links* (connecting form units of one language to lexical units of the same language) and
- *translation links* (connecting lexical units in different languages)



For the translation links, we have made use of the following four parameters:

- conceptual equivalence
- pragmatic contrast
- variant status
- lexicalization status

Conceptual Equivalence: Given the fact that it is hard ever to find true interlingual synonyms, one could of course argue that this interlingual information (i.e. the difference between translation-equivalent items from different languages) could be calculated from the data contained in the monolingual reference lexicons, in particular in their semantics section. However, taking into account, on the one hand, the fact that an explicit and formal system to represent lexical meaning is not available at present and, on the other hand, that, when linking two languages to each other, it is very useful to do so in as economical and powerful as possible a way, we consider it very useful for explicit links to inform about the so-called degree of conceptual equivalence. Let's take the case of *sombra*, in Portuguese. In English, *sombra* is to be translated by two hyponyms, *viz. shade* (focusing on the locative aspect of *sombra*) and *shadow* (focusing on the formal aspect). If one has a set of values such as (see Martin-Tamm 1996, p.683)

- *complete equivalence* (i.e. there is a complete conceptual equivalence between Source Language LU (further on: SLU) and Target Language LU (further on: TLU)

- *hyperonym* (i.e. the TLU is the hyperonym of the SLU)

- *hyponym* (i.e. the TLU(s) is/are (a) hyponym(s) of the SLU)

- *substitution by near equivalent* (e.g. English *barrister* vs. Portuguese *advogado*)

- *related* (e.g. the English series *shine, glimmer, glisten, glow, glitter* and the Portuguese series *brilhar, cintilar, resplandecer*)

then this set of values not only makes the choice between equivalents easier, it also makes the reversal of equivalents much more accurate when moving from A-B to B-A, a fact which is extremely important in the dictionary making process. Compare the case where the hyponym-link between the English LU *inflection* and the Dutch *verbuiging, vervoeging* will be inversed into a hyperonym-link when the reversal takes place. Explicitly stating the relationship here implies that the semantic constraints in the hyponym-linking (viz. <w.r.t. nouns> and <w.r.t. verbs and adjectives>, respectively) will now be transformed into semantic specifications, thus e.g. Dutch *vervoeging* = E. *inflection* <of verbs>.

Pragmatic Contrast: As each Lexical Unit should be accompanied by a specification of its pragmatic value, a pragmatic calculus can be carried out to yield a signaling or pragmatic contrast/similarity in the case of linking two lexical units. So e.g. although from a conceptual point-of-view the English word *bed* and the Dutch words *bed, nest* and *sponde* all refer to the same concept, their pragmatics greatly differ. In both languages *bed* is the neutral term, Dutch *nest* being informal, *sponde* being formal and obsolete. As a rule the pragmatic calculus can be adjusted so for it to yield a blocking between items belonging to different types (marked vs. unmarked), or a blocking only when going in one particular direction, allowing e.g. the direction marked -> unmarked, but not vice versa (an example of such a pragmatic calculus is to be found in Fernandes, C. , 1995).

Variant Status: Although some dictionaries may create the impression that there is a one-to-one correspondence between most items from different languages, one should bear in mind that this is rather the exception than the rule. Good bilingual dictionaries are proof of the fact that, when it comes to comparing languages, one is often dealing with many-to-many relations.

Let's take the case of Portuguese *sombra* again, translated into Dutch by *schaduw, schaduwzijde, lommer, beschutting*. Although one would consider the above translation equivalents to be pointers to the respective LUs in the Dutch

Reference Lexicon, where more information on their degree of equivalence is to be found, still it could be useful and handy to have linkage information available (in this case on the mutual status of the translation possibilities). This would inform us that whereas *schaduw* is an unrestricted, primary translation equivalent, this is not the case of *lommer*, which is both conceptually and pragmatically restricted and therefore considered a secondary variant. Notice that, in order to generate or suggest the so-called cross-linguistic synonyms and pseudo-synonyms, it is useful, not to say necessary, for the Reference Lexicon to contain thesaurus-like data such as synonyms and other lexical relations. So e.g. when *car* gets two translation equivalents in Portuguese (one primary = *carro* and one secondary = *automóvel*), this does not imply that the relation can be reversed as such; indeed *automóvel* will have *car* as its primary, not as its secondary translation equivalent.

Lexicalization Status: By default, the Fus + LUs ocurring in the Reference Lexicons are lexicalized. However, when translating a FU/LU in one language into another one, one sometimes cannot link with a FU/LU from that language, but has to resort to semi-lexicalized objects or non-lexicalized ones, such as descriptions and borrowings. In the latter case there is no link to an existing FU/LU in the Target Language. So e.g. English *haggis* will yield the borrowing *haggis* in French or the description *enchido escocês feito de carne de borrego cozida no próprio estômago do borrego* in Portuguese, the former only sharing the translation status, not that of a French FU/LU, the latter being but an explanation in Portuguese of the English FU/LU.

Whereas the pragmatics are already specified in the monolingual reference lexicon, the lexicalization status has to be specified in an extra data category when "linking" with non-lexicalized objects and/or borrowing or semi-lexicalized objects takes place.

The results of all these translation links at the meaning level are a considerable improvement of the monolingual resources and, consequently, of the quality of the bi- or multilingual dictionaries for the end-users.

A most important and innovative aspect of this LRL is the contribution of a multimedia component to help the lexical description. By making use of the different modalities of knowledge representation (lexicon, text corpora, encyclopedic knowledge, thesauri, galleries of images, drawings, videos and sounds) in an intelligent, cognition-oriented way, one can come to a more complete and dynamic lexicon.

The main relevance for the introduction of a multimedia component in this project of LRLs is justified by the fact these different modes of knowledge representation have been up to now rather poorly explored in lexicographic products and, above all, with an extreme lack of fundamental scientific research.

Just by way of exemplification, let us state that most of the CD-ROM dictionaries and encyclopaedias now available on the market do not differ too much from their printed versions in terms of the information contents. What happens in most cases is that images and sounds (that have often been previously in the publisher's archives and therefore not produced or analysed by lexicographers) are added to the text, but are used very much at random, mainly serving marketing and layout purposes. The main alterations produced in those CD-ROMs obviously have to do with the different means of access to the information, but above all they have to do with surface questions like the use of more attractive layouts rather than with contents questions. In other words, the state-of-the-art of the "new" dictionaries is quite disappointing in lexicographical terms. Where one would expect a real advance, one finds a set of "decoration" features that contribute to the products' appearance, but rarely to a radical change in the **concept** of what a dictionary is or should be.

We believe that the use of the different media for knowledge representation at our disposal to date can be of great help to language learners when confronted with difficulties of different kinds, be it of meaning differentiation in polyssemic words, insufficiency of pure linguistic means to explain word meanings, sense ambiguity or culture-bound units, for instance.

The background idea is that a database for LRLs can be composed of various modules that correspond to the above mentioned sources of knowledge: the lexicon, definitions, text corpora, encyclopaedic information, thesauri and different galleries of images, drawings, videos/animations and sounds. These modules will be activated in an intelligent and dynamic way (through the use of slots and fillers of pre-defined frames) according to the lexical item that is being accessed. For example, if a user wants to look up a verb like "slide" (which is stored in the cluster of "verbs of movement"), the first source of information to be activated should be a video or a photo illustrating that action, then a definition can be provided and finally some examples from the corpus can be shown. A totally different order should be followed in the case of a word like "law", where an ostensive illustration is impossible and where a definition and some examples are the most appropriate modes. For a word like "shrill", where a definition of the type "a high and unpleasant sound" is clearly not illustrative and specific enough, a reproduction of a shrill voice would probably be the best mode of representation, before the user is led to definitions or to contextual examples.

Along the research process, we have decided to give priority to the parts of speech that are not usually ostensively illustrated, given the fact that the existing lexicographical products (both printed or in CD-ROM format) practically always and only use illustrations for the nouns category, i.e., mainly for concrete objects, animals, plants, places and famous personalities.

Taking this into account, our LRL will provide as much multimedia material as possible for the following word clusters:

- adjectives in contrast sets (e.g. *convex/concave; centrifugal/centripetal*): using photos
- verbs (mainly motion ones): using videos
- prepositions (in their basic spatial uses): using videos and drawings
- verbs, adjectives and nouns related to sound: using sound recordings
- words that are commonly used in an abstract or metaphorical way (e.g. *hold: she has never held ministerial office*) by showing their basic or physical meaning: using photos and videos
- polysemous nouns (bank, arm, etc.): using photos and/or drawings
- taxonomies (e.g. *quadrilateral: square->rectangle->parallelogram*): using drawings
- partonomies (e.g. *bicycle: saddle, handlebars, wheels, etc.*): using photos and/or drawings
- paradigms (e.g. *livestock: adult, young, infant*): using photos and/or drawings
- cycles (e.g. *morning, afternoon, evening, night*): using photos and videos.

References

ABRAMSON, H. et. al. 1996, "Multimedia, Multilingual Hiperdictionaries: A Japanese<-> English Example", California University, to be published.

FERNANDES, C. M. 1995, *Lexicografia Computacional: um contributo para novos dicionários bilingues*, Dissertação de Mestrado, Universidade Nova de Lisboa, FCSH.

FERNANDES, C. M. 1996, "Multimedia dictionaries: the interaction between word and image", Vrije Universiteit, Amsterdam.

FILLMORE, C. 1977, "Scenes-and-Frames Semantics" in *Linguistic Structures Processing*, ed. by A. Zampolli, Amsterdam.

JOHNSON, M. 1987, *The Body in the Mind*, The University of Chicago Press.

LAKOFF, G. and JOHNSON, M. 1980, *Metaphors we live by*, The University of Chicago Press.

LAKOFF, G. 1987, *Women, Fire and Dangerous Things*, The University of Chicago Press.

MARTIN, W. 1991, "On the dynamic organization of computer lexicons" in *Perspectives on the English Lexicon*, ed. by S. Granger, Louvain-la-Neuve.

MARTIN, W. 1994, "Knowledge representation schemata and dictionary definitions" in *Perspectives on English studies in honour of Professor E. Vorlat*, ed. by Carlon, K. e.a., Leuven: Peeters.

MARTIN, W. and TAMM, A.(1996), *OMBI: an editor for constructing reversible lexical databases* in M. Gellerstam e.a. (ed.), Euralex'96 Proceedings, I-II, Göteborg University, 1996

MILLER, G. et. al.1990, "Introduction to WordNet: an on-line lexical database" in *International Journal of Lexicography*, Vol.3 N° 4, pp. 235-244, Oxford University Press.

MILLER, G. 1986, "Dictionaries in the mind" in *Language and Cognitive Processes*, 1: 171-85.

MINSKY, M. 1975, "A framework for representing knowledge" in *The Psychology of computer vision*, ed. by P.H. Winston, pp. 211-277, New York.

MINSKY, M. 1977, "Frame system theory" in *Thinking: readings in Cognitive Science*. ed. by P.N. Johnson-Laird and P.C. Wason, Cambridge: CUP.

TAYLOR, J. 1989, *Linguistic categorization: prototypes in linguistic theory*. Oxford: Clarendon.

WEGNER, I. 1985, *Frame-Theorie in der Lexicographie*, Tübingen: Niemeyer Verlag.