INDICE

MoLeX: TOMORROW'S COMPUTATIONAL DICTIONARY TODAY

Elena Bárcena
Tim Read
Universidad de Granada

This paper presents MoLeX (Multilingual Onomasiological Lexicon), a prototype which was designed and implemented to illustrate a novel type of on-line lexical reference system that overcomes some of the shortcomings in previous lexicographical work. The paper is divided into two parts. The first part describes the linguistic theory behind the dictionary, which is based on Martín Mingorance's original Functional-Lexematic Model. A number of features are emphasised, such as the complete and economically arranged information within each lexical entry and the hierarchical and onomasiological organisation of the whole lexicon (which has been devised in a bottom-up manner from a selection of existing dictionaries). In the second part, the design of the automatic version of the dictionary is introduced and exemplified in MoLeX. It can be used either mono- or multilingually via an interface which is independent from the database of lexical knowledge. The interface is flexible and portable, and has been written in Java, which makes it accessible both locally, or remotely across an intranet or the Internet via the World Wide Web.

1. INTRODUCTION

This paper presents the design of MoLeX (Multilingual Onomasiological Lexicon), the prototype of an on-line lexical reference system which was recently developed by our research team.[1] MoLeX is the result of further development and implementation of Martín Mingorance's Functional-Lexematic Model (FLM) (1984) for the description of the core vocabulary of natural languages. Major design features of MoLeX include the complete and economically arranged information within each lexical entry and its global hierarchical and onomasiological organisation, which has been devised in a bottom-up manner from a selection of existing dictionaries. Thanks to its compositional design, MoLeX avoids circularity, reveals significant intra- and cross-linguistic generalisations, and its computational version can be used either mono-, bi-, or multilingually. This is achieved via an interface which is independent from the lexical database. The interface is flexible and portable, and has been written in Java, which makes it accessible both locally or via the World Wide Web (henceforth WWW): http: //www.ugr.es/~tread/molex.html (with a Java enabled browser such as Netscape Navigator 4 which can be obtained from http: //www.netscape.com). At present MoLeX covers Spanish and English verbs, but new languages and lexical categories are under development.

The standard type of dictionary organisation is semasiological (i.e., according to form), of which alphabetical order is undoubtedly the most common. Important informa-

[1] This work has been done as part of the project «Development of a lexical logic for computer-assisted translation from a multifunctional and reusable English-Spanish-French-German lexical database» (PB94-0437), which is funded by the Spanish Ministry of Education and Science.

tion about the global organisation of the vocabulary of the language is, however, inevitably lost. This fact and the goal to produce a dictionary which is psychologically adequate led a number of lexicographers to develop onomasiologically organised lexical resources (i.e., according to conceptual meaning), the most common of which are thesauri. Thesauri, however, have a considerable amount of redundancy and provide little or no information about the combinatorial and selectional properties of words. Their organisation is psychologically motivated to some extent, but their top-down structure is considered to be rather ad hoc.

According to Martín Mingorance, a dictionary should reflect the organisation of the human mental lexicon and account for the way in which individual languages lexicalise conceptual knowledge. Psycholinguists like Apresjan (1993) claim that there is empirical evidence about both the nature of semantic relations, which are not arbitrary but real mental functions, and the hierarchical structure of the human psyche, with semantically related words stored near to each other in the mental lexicon. In accordance with this view, Martín Mingorance outlined the design of a relational lexicon-based on meaning structure, where lexemes are organised hierarchically in a bottom-up manner according to their common and differential semantic components. Not surprisingly, the inventory and internal structure of lexical fields and the distribution of lexemes within them which have been arrived at through this method sometimes differ from those normally found in traditional thesauri. He also believed that linguistic competence largely depends on the information obtained from the mental lexicon, which implies the existence of complex linguistic (e.g., morphosyntactic, semantic, pragmatic) knowledge attached to its lexical items. Therefore, he undertook a study of the type of paradigmatic and syntagmatic information which would be necessary to include in the entries of a dictionary in order to account for successful linguistic production and comprehension, and how to do this economically. Martín Mingorance incorporated cognitive issues like these into the elaboration of his lexical theory which he called FLM.

2. Functional-Lexematic Model

FLM is based on Martín Mingorance's integration of two linguistic theories, namely Dik's Functional Grammar (FG) (1978) and Coseriu's Lexematic Theory (LT) (1977), in order to obtain both a syntagmatic and a paradigmatic axis which describe lexical combination and selection respectively. The FG-based syntagmatic axis uses predicate frames as integrated formulae which specify the combination patterns that govern the predicates of a language. On the paradigmatic axis, lexemes are arranged onomasiologically within semantic fields following the dictates of LT. Microstructurally, lexical entries are characterised as complex units of syntactic, semantic, and pragmatic information. On the macrostructural level, lexical entries are interconnected semantically in a number of ways forming complex hierarchies. The resulting integrated model is intended to allow the lexicographer to build a hierarchical network of semantically related lexemes within general areas of meaning, with highly informative and economical definitions.

FG and LT are not only complementary but also compatible from a number of theoretical and practical viewpoints. As Dik (1978: 46) says: «Although the view of lexical analysis fits in nicely with the model of FG, the assumptions embodied in it do not necessarily follow from this model. That is, FG would also be compatible with other conceptions of lexical meaning definition». For example, FG is teleologically functional because it regards language as an instrument of verbal interaction, not as an abstract system autonomous of real usage. LT is structurally functional in that language is seen to be determined by a system of functional oppositions. Furthermore, the FG principles underlying the structure of meaning definitions and the procedure of stepwise lexical decomposition for definitions have many points of convergence with the principles of lexical field theory and factorisation of LT, as subsumed in FLM (Mairal, 1993).

FG was adopted by Martín Mingorance for a number of reasons, including its psychological adequacy and its integration in a theory of verbal communication. FG is lexicon-driven, providing a complete account of the combinability of lexical elements in the linear sequence, stated in functional and abstract terms. Lexical units are regarded as structured representations, and encoded in the form of predicate frames, which encapsulate the following information:

(a) form of the predicate: the orthographical and phonological representation of the predicate;

(b) syntactic category: the part-of-speech label; namely, verb, noun, or adjective;

(c) quantitative valency: the number of arguments subcategorised by a given predicate;

(d) qualitative valency: the semantic role of the arguments in the state-of-affairs designated by the predication;

(e) selection restrictions: the features which specify the nature of the arguments which can appear in complement and subject positions;

(f) meaning definition.

Meaning definitions follow the strategy known as stepwise lexical decomposition and, similarly to sentential underlying predications, are stated in terms of predicates which exist in the language. Due to the complexity of devising a finite set of universal units of semantic description (e.g., binary features, abstract primitives, atomic predicates) as reported by numerous researchers, the meaning components used in FG are natural language phrases, although there are several constraints in their expression. Apresjan (1993, p.86) writes: «In the majority of languages there are semantic primitives, or rather, *near primitives*, to describe the basic concept of each system. The defining predicates occurring in meaning definitions are lexical items of the object language» (our emphasis).

It should be noted that in FLM there is considerable refinement of the content of predicate frames as stated by standard FG. For example, for each given predicate there is a list of one or more preferred predicate frames, which accounts for syntactic flexibility and avoids the rigid problematic distinction between arguments and satellites in FG. Also, the descriptiveness of predicate frames is enriched with further information; for instance, the syntagmatic description of verbs includes collocational information and syntactic functions which are not present in standard FG.

Martín Mingorance incorporated LT into his model in order to expand the knowledge in the lexicon beyond the level of individual entries reflecting its global relational structure, and hence to build the dictionary of a language as a hierarchical network of semantically connected lexemes. LT contains the following fundamental concepts:

(a) lexical field: the set of all lexical units which share an explicitly distinguished non-trivial semantic component;

(b) dimension: a subdivision of a lexical field which falls halfway between minimal groups of lexemes and the lexical field proper; a viewpoint or articulation of the content of the lexical field directly derived from the definitional structure of lexical units;

(c) archilexeme: the nuclear word or definiens in terms of which all the words in the lexical field are defined;

(d) seme: the minimal feature in the meaning definition of lexemes which differentiates them from each other within a dimension;

(e) classeme: a content feature by which a lexical class is determined; it can operate throughout the lexicon in different dimensions and lexical fields.

With this relational model, it is possible to capture the semantic hierarchical structures of whole lexical fields, the relationships (hyponymy, synonymy, antonymy) between their items, and the subtle semantic nuances of such items. Furthermore, following the compositional method of definition and the principle of feature inheritance, the hyponyms of a given archilexeme become the definiens of other words at more specific levels of the hierarchy, and so on. The 'near primitive' components which form meaning definitions are inherited by all the lexemes lower in the hierarchy, so that definitions are clearly expressed in such simple terms as the combination of the immediate superordinate plus their nuclear meaning via one or more components. This economical method prevents circularity in the definitions and reveals significant generalisations about the structural semantic organisation of the lexicon of a language (Bárcena et al., 1997).

The development of a framework which organises and describes predicates in the lexicon following lexematic principles leads to a deeper understanding of individual languages and the complex relationship between syntax and semantics. Regarding the semantic structure of lexica, for example, a number of principles have been established about the recurrence of dimensions across lexical fields (Faber, 1994). Furthermore, various syntactic aspects of words are tied to their meaning. The close relationship between syntax and semantics has been observed by authors like Dixon (1991, p.75): «The lexical words of a language can be grouped into a number of semantic types, each of which has a common meaning component and a typical set of grammatical properties». Many of these properties, like classemes, are recurrent in other domains and lexical fields. Martín Mingorance developed the notion of predicate schema as a kind of extended predicate frame which captures the syntactic, semantic and pragmatic information shared by all the lexemes in each dimension. This is, once again, crucial both to organise the entries more economically, since only new syntagmatic information needs to be stated for each predicate lower in the hierarchy, and also to capture significant syntactic and semantic generalisations within semantically related lexical items.

The methodology for building the FLM dictionary is data-driven and bottom-up. The primary source of data is a collection of widely used monolingual dictionaries. In order to arrive at the definition of a given word, all the dictionaries are consulted and the corresponding lexical entries are analysed. The relevant meaning components for the identification of the word, which have been taken from the existing definitions, are selected and put under headings such as nuclear meaning, direct object, adverbial complementation, and pragmatic information. A new structured definition is then built by reformulating the selected information according to certain expression rules. Definitions are composed from left to right, starting from the definiens and adding increasingly specific features on the right, which differentiate the word from the preceding members of the hierarchy. The method of assignment of a word to a lexical field is the prototypical method, according to which there is a set of clear cases or prototypes which serve as points of reference. Classemes and dimensions are established on the basis of the presence of recurrent patterns as the lexical hierarchies are being built. The inventory of lexical fields within the vocabulary of the language is obtained. After having elaborated the definitional structure of 10, 000 English verbs, it has been found that words fall into the following basic domains: EXISTENCE, MOVEMENT, POSITION, CHANGE, POSSESSION, PERCEPTION (including stimulus verbs such as LIGHT and SOUND), EMOTION, COGNITION, SPEECH, and GENERAL ACTION (composed of subgroups such as verbs of CONSUMPTION, COMPETITION, CONTACT, USE, etc.). The lexical fields of the different languages are compared and cross-linguistic links are created between equivalent lexical entries. These links are not always one-to-one so when a lexical gap is found in a given language it must be filled by inserting a periphrastic expression. Furthermore, an indexing system needs to be elaborated to cater for the different senses of ambiguous words.

3. MoLeX: THE COMPUTATIONAL PROTOTYPE

As is common practice today in Lexicography and generally in Linguistics, a computational prototype, MoLeX, was developed to test and evaluate FLM. Furthermore, it was built as a step toward the implementation of the full scale dictionary, which is underway. Consulting paper-based lexical resources is often a tedious and cumbersome task. As Hutchins & Somers (1992) comment, it has been estimated that professional translators spend over 60% of their time consulting such resources. For a long time there has been great interest in making the task of lexical consultation faster and more user-friendly. It did not require much imagination to realise the role that computers could play in this process (not only in the provision of lexical resources, but also, for example, word processors and grammar and spelling checkers). Lexicographical systems have many advantages over paper-based versions including the following:

(a) fast and flexible look up: data from automatic dictionaries can be accessed via keywords (or parts thereof) other than the exact entries themselves, or even combinations of keywords;

(b) large storage capacity: as memory and hard disk sizes grew and optical storage media (such as compact disks) became available, on-line lexica were increasingly able to offer much larger linguistic coverage than their paper-based counterparts;

(c) multilinguality: paper-based dictionaries are normally either mono- or bilingual and cover a small collection of languages. On-line sources, however, offer multiple language combinations;

(d) fast and flexible use: computational dictionaries can be used on their own, or as part of integrated environments such as the translator's workbench;

(e) ease of modification, updating, and customisation: new entries can be easily added, either as part of the main dictionary or as separate specialised sub-dictionaries;

(f) test-bed: in a research context, computerised tools provide a practical way to investigate languages and the functionality and implications of linguistic theories.

Furthermore, Computational Lexicography has found a novel and dynamic application environment in the Internet. The evolution of the Internet has been (relatively speaking) quite fast. It was started in 1969 by the American government for use in defence research projects. By 1983 there were five hundred systems connected and by the end of 1994, more than four million. The design of a new transfer protocol HTTP (Hypertext Transfer Protocol) by the CERN research laboratory in 1990 improved significantly both the quantity and the type of information that was being transferred across the Internet. Such transfer (structured by a Hypertext Mark-up Language [HTML]) using HTTP gave rise to the WWW. The birth of the WWW has enabled on-line linguistic tools such as glossaries and dictionaries to be implemented and accessed from computers in many different countries.

The WWW enables a large number of people to simultaneously access the same tool, although the currently small number of private connections to the Internet limits the number of users. Another characteristic of the dictionaries present on the Internet is the ease with which they can be modified, since there is only a single copy of the dictionary located on the server, and not a copy installed on each user's computer. Furthermore, the centralisation of linguistic tools like dictionaries reduces the amount of space which each user has to reserve for personal copies. One of the most important issues in the use of such centralised tools is the quality and cost of the Internet connection. If someone is expected to make use of WWW based linguistic tools, they must have access to fast, reliable, and cheap Internet connections. It should be noted that there are many places where this is not the case, but the situation is generally improving.

One serious problem that these dictionaries (and other Natural Language Processing [NLP] tools) have had until very recently is the lack of sophistication present in their interfaces. On-line dictionaries written in HTML are sufficiently sophisticated to enable the user to select an entry from a list (which can only be scrolled through) to access its definition or foreign language equivalent. However, such a dictionary cannot provide the functionality of the interfaces found in dictionaries on desktop computers (e.g., the ability to automatically search for specific entries; Heslop & Budnick, 1996). The arrival of the Common Gateway Interface (CGI) enabled HTML documents to be connected to server resident programs which can perform such additional operations. This technology is still insufficient to provide the types of sophisticated and flexible interfaces which would be needed to build general on-line NLP tools.

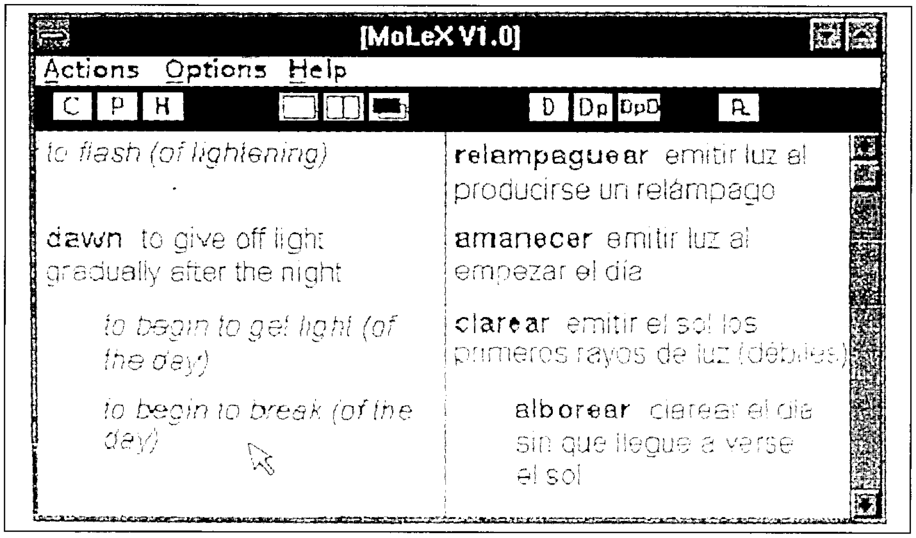
The Java programming language represents the core of a technology which can be used to build WWW based programs (applets) that provide the same type of functionality present in conventional computer based tools. Java was designed and released officially by Sun Microsystems in 1995 as a simple, robust, dynamic, multi-threaded, object oriented programming language, which can be used to produce programs that run independently from both the hardware and the operating system of the target computer (due to the Java virtual machine) (Naughton, 1996). The Java applets run within HTML documents independently of the location of the program in relation to the user's machine (locally or on a machine on the other side of the world!).

Java has consequently been adopted by the design team for the production of a prototype of an automatic dictionary (MoLeX) based on FLM (Read et al., 1997), which at present covers English and Spanish verbs. The development of a pilot system is a common practice today in the area of Lexicography and Linguistics in general. Furthermore, its construction has been used to evaluate FLM and is a first step toward a complete implementation of the dictionary, which is underway.

Like the majority of programs with a visual interface, MoLeX is an event driven system in which (following the initial load and set-up phase) the actions of the user determine the information that is presented on the screen. The figure below illustrates the MoLeX interface (v.1.0) showing the paradigmatic axis. The interface includes three menus and an icon bar (which duplicates the most important functions of the menus). The linguistic information contained in the dictionary can be displayed in three ways: the dimensions of a given lexical field, the dimensions together with their lexemes, and the dimensions together with their lexemes and definitions.

The selection of a dimension or word (by clicking on it with the mouse) retrieves its equivalent(s) in the other language. A selected entry can be copied or printed for future use, and a basic history list enables the user to return to entries selected earlier in the session. Various properties of the interface can be changed, such as the interface language, colour, and font properties. It should be noted that the linguistic information in the two languages can be presented to the user in three different ways: firstly, as two lists presented side by side, aligned and connected to the same vertical scrollbar (as in the figure below). Secondly, as two non-connected lists, presented side by side, with separate scrollbars. Thirdly, as a single list which spans the entire window and contains its own scrollbar (in which case the user can swap between windows).

The interface allows the user to consult the different meanings of a given word. There are two types of ambiguity in this system: ambiguity in the source language and in the target language. In the first type, the same word can appear in more than one dimension or more than once under the same dimension. The user is offered all the possible interpretations of the ambiguous word and then prompted to select the one required. In the second type of ambiguity, a word has more than one equivalent in the target language. The distinct occurrences of the word can be distinguished in terms of the corresponding dimensions or simply by its translation equivalents. It should be noted that the format of the database provides information related to the descriptive model. For example, the relations of hyponymy between the different dimensions and between the lexemes in each dimension are represented visually by indenting the left margin (as can be seen in the example below).



to flash (of lighlening)

dawn to give off light gradually after the night

to begin to get light (of the day)

to begin to break (of the day)

relampaguear emitir luz al producirse un relámpago

amanecer emitir luz al empezar el día

clarear emitir el sol los primeros rayos de luz (débiles)

alborear clarear el día sin que llegue a verse el sol

4. CONCLUSION

This article presented an on-line lexical reference system for the description of the basic lexicon of natural languages, which follows a new dictionary model based upon FLM. This is a theory based upon cognitive principles which presents the dictionary as the representation of the mental lexicon of the speaker. FLM has been designed to construct both an onomasiological description of the vocabulary of a language where the predicates are grouped into semantic classes or domains, and a complete description of the predicates. FLM integrates FG together with LT to account for the syntagmatic and paradigmatic relations in the lexicon, which are based upon the complementary principles of combination and selection present in definitions. The knowledge in the lexicon is extended beyond the level of the individual entries in order to reflect its global relational structure, which allows the lexicon to be designed as a hierachical network of semantically connected lexemes. This network is dynamic and open to the progressive incorporation of further entries.

MoLeX has been designed in two parts: a database of lexical knowledge and an interface written in Java. The onomasiological structure of the information in the fields and tables within the database permits the linguists working on MoLeX to access and modify it easily, without needing to take note of any complex program related data structures (e.g., the data files written for Prolog based systems [cf., Dik, 1992]). The interface to the lexicon is formed by an HTML document that contains a Java applet which controls access to the information held within the database. While in some respects the interface depends on the WWW browser used (e.g. Netscape), its functional aspects are very similar to most WWW pages, independently of whether the lexicon is located on the same computer as the browser or on another machine. The sophistication of the interface comes from the Java applet which accesses the database and undertakes the linguistic processing. Furthermore, the combination of HTML and Java implies that the lexicon can be used on a wide range of hardware platforms. All the help and documentation is available via on-line HTML documents, for which the user does not have to learn a new way to interact with the information. Finally, the inclusion of an email command enables the users to ask the designers questions, and the designers, to obtain feedback from users.

Bibliography

Apresjan, J. D. 1993: Systemic lexicography as a basis of dictionary-making. *Dictionaries: Journal of the Dictionary Society of North America* 14.

Bárcena, E., T. Read & P. Faber. 1997: El diccionario del traductor del mañana. Presented at the I Congreso Internacional de Estudios de Traducción (Universidad de La Coruña) and in submission to *Revista de la Facultad de Filología.*

Coseriu, E. 1977: *Principios de semántica estructural.* Gredos. Madrid.

Dik, S. C. 1978: *Functional Grammar.* Foris. Dordrecht.

Dik, S. C. 1992: *Functional Grammar in Prolog. An Integrated Implementation for English, French, and Dutch.* Mouton de Gruyter. Berlin.

Dixon, R. M. W. 1991: *A new approach to English grammar on semantic principles.* Clarendon. Oxford.

Faber, P. 1994: The semantic architecture of the lexicon. > K. Jensen & H. Pedersen eds. *Proceedings from the VI International Symposium on Lexicography.* Niemeyer. Tübingen.

Heslop, B. & L. Budnick 1996: *Publicar con HTML en Internet.* Paraninfo. Madrid.

Hutchins, W. J. & H. L. Somers 1992: *An Introduction to Machine Translation.* Cambridge University Press, Cambridge.

Mairal, R. 1993: *Complementation patterns of cognitive, physical, perception and speech act verbs in English.* Ph.D. Thesis. Universidad de Zaragoza, Zaragoza.

Martín Mingorance, L. 1984: Lexical Fields and Stepwise Lexical Decomposition in a Contrastive English-Spanish Verb Valency Dictionary . > R. R. K. Hartmann ed., *LEX eter 83 Proceedings: Papers from the International Conference on Lexicography at Exeter*. Niemeyer. Tübingen.

Norton, P. 1996: *The Java Handbook*. Macgraw Hill, New York.

Read, T., E. Bárcena & P. Faber 1997: Java and its role in Natural Language Processing and Machine Translation. *Proceedings of MT Summit VI*. Logos Corporation. San Diego.

INDICE