The MuLeXFoR Database: Representing Word-Formation Processes in a Multilingual Lexicographic Environment

Bruno Cartoni[1], Marie-Aude Lefer[2]

[1] LIMSI-CNRS,
BP 133, 91403 Orsay Cedex, France

[2] Centre for English Corpus Linguistics/Université catholique de Louvain
Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgium

cartonib@gmail.com, marie-aude.lefer@uclouvain.be

Abstract

This paper introduces a new lexicographic resource, the MuLeXFoR database, which aims to present word-formation processes in a multilingual environment. Morphological items represent a real challenge for lexicography, especially for the development of multilingual tools. The database introduced in this paper tries to take advantage of recent advances in electronic implementation and morphological theory. Word-formation is presented as a set of multilingual rules that users can access via different indexes (affixes, rules and constructed words). MuLeXFoR entries contain, among other things, detailed descriptions of morphological constraints and productivity notes, which are sorely lacking in currently available tools such as bilingual dictionaries.

1. Introduction

Morphological items and processes pose major challenges for lexicographic work, especially with respect to bilingual and multilingual resources. Affixes usually display several meanings and thereby take part in different word-formation processes. It is therefore difficult to provide enough information to help dictionary users understand the meaning(s) of an affix and the ways it is used to coin new words. In fact, paper dictionaries often fail to achieve this goal. The MuLeXFoR database tries to take advantage of recent advances in morphological description and in electronic multi-access database systems. The prototype so far centres around productive prefixation in English, French and Italian.

The paper is structured as follows. It first gives the reasons for presenting morphological items in a multilingual lexicographic resource. It then briefly introduces the theoretical framework on which the multilingual approach is based. Third, the database architecture is described, with special emphasis on the multiple access points that were adopted to help users understand the multi-faceted nature of morphological processes. Finally, issues of data collection and implementation are briefly discussed, as well as ongoing and future developments.

2. Context: Morphological Processes in Dictionaries

Many (bilingual or monolingual) dictionaries include morphological items in their lists of entries, usually with the purpose of providing information about how to interpret and produce new words. As regards bilingual dictionaries, this type of morphological information is intended to help users understand, translate (and coin) new words in the target language.

The representation of morphological processes in monolingual and bilingual dictionaries has often been criticised in lexicographic studies (Prcic 1999; Dardano et al. 2006; ten Hacken et al. 2006; Cartoni 2008a; Lefer 2009). Importantly, these studies have put forward the inadequacy of relying solely on affix representation, which is how morphological items have been included in (paper) dictionaries so far. Two semantic issues need to be addressed. First, prefixes frequently display a range of possible meanings, such as English *pro*, which can convey both "hierarchy" (e.g. *proconsul*) and "support" (e.g. *pro-independence*). Second, meanings are often conveyed by several prefixes (e.g. "unspecified plurality" and Italian *multi*, *pluri* and *poli*). These two phenomena represent a serious challenge, especially in multilingual tools, and require the adoption of a sound theoretical framework.

3. Theoretical Framework: the Lexematic Approach

The lexematic approach to morphology (see Fradin 2003 for a summary of the most recent studies in this field) considers affixes as the formal components of Lexeme-Formation Rules (hereafter LFRs) which entail other constructional operations (e.g. word category change, semantic operation) and which, most importantly, are semantically-driven.

Some monolingual tools rely on this rule-based approach to rationalise morphological information (e.g. Bernal's DSVC 'database for Catalan affixes'; see Bernal and DeCesaris 2008). The present project is largely inspired by Bernal's monolingual database.

Lexematic morphology proved to be extremely useful to formalise multilingual LFRs that match equivalent constructional processes in different languages. For example, one can formalise a "reiterativity" LFR that

creates verbs from verbs (LFR_reiter(v→v)) and represents the various affixes that are used cross-linguistically to express this meaning (*ri* in Italian, *ré* in French, *re* in English). The semantics of the rule is used as the pivot of the translation process (*ri*, *ré* and *re* can be theoretically considered as the surface forms of one single cross-linguistically valid LFR).

A further advantage of this approach applies to cases of synonymy, where one rule represents several affixes. For example, the "unspecified plurality" LFR consists of three prefixes in Italian and French and two prefixes in English (IT: *multi, pluri, poli*; FR: *multi, pluri, poly*; EN: *multi, poly*). In these cases, monolingual constraints can be specified in the database entries to help users select the appropriate affix.

Another interesting aspect of the lexematic approach concerns the coinage of prefixed relational adjectives. These adjectives are derived from suffixed nouns ("Xsfx", where X is the nominal base). The semantic operation of the prefixation rule applies to the base noun. For instance, this rule implies that *multidimensional* can be paraphrased as "with many dimensions" (see Fradin 2008 for a complete description of this phenomenon). This is the reason why the word category change is represented as (n > a).

The lexematic framework provides formalisation methods and theoretical tools which are particularly useful for presenting word-formation in multilingual lexicographic tools systematically and rigorously.

4. MuLeXFoR: General Architecture

The MuLeXFoR project aims to present multilingual LFRs (as described in Section 2) in a user-friendly interface. The system is based on unified morphological rules, which are core to the database. English, French and Italian prefixation processes have currently been implemented in the system[1].

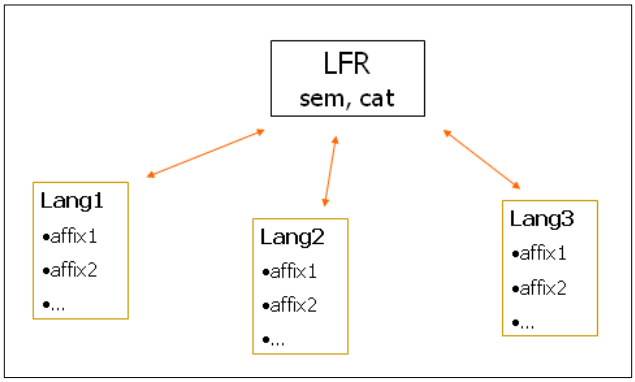

Figure 1: 2-level architecture

As shown in Figure 1, the morphological processes (or Lexeme-Formation Rules) have surface representations in each language (e.g. affixes and other morphological processes such as conversion or compounding, which have not been included in the prototype database yet).

In terms of meta-lexical information (e.g. instructions on affix use and productivity), specific fields are provided at the affix level (for monolingual specific information) and at the rule level (for general information).

As regards the implementation of the tool, the use of a multi-access and dynamic database enables users to access morphological information via different modes and languages. First, users can browse the database via semantic labels, thus accessing whole multilingual LFRs and their respective affixes and constraints (see Section 3.2). This access mode obviously requires a high level of morphological competence. This is why users can also browse the database via the affix index for each implemented language, as described in Section 3.1.

4.1 Affix Browsing

Users can select the affix they wish to look up in the affix index, as is the case in any traditional dictionary. The originality of our approach lies in the fact that when clicking on an affix, users have access to the rules that the affix takes part in, a complete description of each rule and the corresponding equivalent affixes in the other target languages. For example, users who wish to know how to express English *multi* in French can first select the English prefix *multi* in the affix index. MuLeXFoR then provides the rule(s) that involve(s) this English prefix (in this case, "unspecified plurality" to coin adjectives from nouns (n>a) and nouns from nouns (n>n)). When clicking on one of these rules, users get a comprehensive description of the multilingual rule, including the equivalent affixes in Italian and French (*multi*, *pluri*, *poli*/*poly*), their usage restrictions, and examples. This is illustrated in Figure 2 (see Appendix) for prefix *multi* within the "unspecified plurality (n>a)" rule.

As can be seen from Figure 2, two usage notes are provided. The first one is rather general and concerns the use of the prefix *poly/poli* in the three languages, while the second is specific to French and identifies other non-morphological ways to coin the same meaning as the prefixes (here a prepositional phrase).

4.2 Rule Browsing

The database can also be browsed via specific rules. Figure 3 (see Appendix) illustrates the "Location space – Above" rule which coins adjectives from nouns.

Once we click on the rule name in the menu panel, the selected rule subsequently appears in the main panel. This provides various types of information (affixes, morphographemic information, etc.). It is quite obvious that this type of browsing is not easy for non-expert users. LFRs are currently presented in the language of the platform (English) but we plan to localise the interface (at least in French and Italian). We are also presently

[1] MuLeXFoR is available on the web. Please contact the first author for login and access information.

developing an interface specifically designed for second language learners and trainee translators where instructions, menu names and rule names are carefully adapted to suit learners' needs (see Cartoni & Lefer 2010).

4.3 Other Possible Access Points

The database can also be accessed via the lexical index, which consists of all the examples provided in the LFRs for the three languages. Interestingly, this lexical index could provide a link between the constructed lexemes of a bilingual dictionary and the MuLeXFoR database. It is important to note that there is also room for improvement regarding this aspect of the database. We envisage adding an automatic morphological analyser component (see e.g. Derif; Namer 2002). Complex words that are not included in the database (e.g. neologisms) could be automatically analysed and subsequently matched to the corresponding rule. Needless to say, this feature would depend heavily on the efficiency of the morphological analyser used and on the exhaustiveness of the database.

5. Data Acquisition Issues

As in any lexicographic work, data collection (to feed the database) is a thorny issue. In morphological resources such as MuLeXFoR, two main types of knowledge need to be acquired and formalised. On the one hand, the multilingual rules (i.e. cross-linguistically shared syntactic and semantic operations) have to be singled out. On the other, productive affixes (and other productive morphological processes) corresponding to these rules need to be identified in each language.

The first implementation step was largely inspired by the linguistic literature that provides abstract – and hence cross-linguistically valid or even universal – semantic descriptions of morphological processes. As argued in Szymanek's (1988) study, morphological processes are closely related to basic cognitive notions, such as movement, modality, evaluation, etc. By examining various semantic descriptions of prefixation in different languages (e.g. Montermini 2002, Iacobini 2004 for Italian; Amiot and Montermini 2009 for French), a rather exhaustive set of rules was identified (see Cartoni 2008b for further details).

The semantic categories implemented in MuLeXFoR currently focus on prefixation and, to a lesser extent, conversion. Even though suffixation is usually said to be more abstract and semantically less specified than prefixation, a similar approach could be applied to suffixes.

Corpus-based methods and tools were used in the second stage where we aimed to determine which prefixes contribute to which rule(s) in the three languages investigated. We drew from the results of a detailed study on word-formation which focused on machine translation from Italian into French. This study heavily relied on corpus data (*La Repubblica Corpus*; Baroni et al. 2004) (see Cartoni 2008b). The English data along with additional French data were collected within the framework of a corpus-based contrastive study of English and French prefixation across genres (press editorials, novels and scientific articles) and academic disciplines (c. 100 prefixes in each language were investigated; see Lefer 2009). Both corpus-based studies made it possible to single out productive prefixes in the three languages investigated, together with authentic examples of neologisms formed with these prefixes.

MuLeXFoR relies on these two data-intensive studies. It currently contains more than 60 multilingual LFRs and 50 productive prefixes in French, Italian and English. Further data acquisition methods are presently under investigation to increase the coverage of the database.

6. Assessment Issues

As stated above, the database is a prototype and a number of assessment issues still need to be addressed. Assessment of the database is planned, mainly in terms of users' needs and expectations. Originally, the MuLeXFoR database did not target an audience in particular. However, we soon realised that the labels used in the interface were too opaque for non-expert users such as second language learners or trainee translators, who might struggle with linguistic terminology. The user-oriented assessment of the tool will therefore focus on these non-expert users (e.g. in terms of the comprehensibility of the notions used in the menus). Browsability and access to information will also be evaluated.

In addition to user-oriented evaluation, MuLeXFoR's interoperability with existing lexicographic tools will be assessed.

7. Conclusion and Further Work

In addition to the obvious extension of the resource to other affixes and to other languages, the inclusion of conversion (where no surface forms are implied) and other morphological items (e.g. neoclassical constituents such as *paleo*, *bio*, *eco*) will be examined.

Although MuLeXFoR is still under development, we hope that the framework presented here will contribute to the improvement of the representation of morphological items in multilingual databases and tools.

8. Acknowledgments

We would like to thank the anonymous LREC reviewers for their insightful comments.

References

Baroni, M.,Bernardini S., Comastri F., Piccioni L., Volpi A., Aston G. and Mazzoleni M. (2004) Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In proceedings of LREC 2004, Lisbon pp. 1771-1774

Bernal, E. and DeCesaris J. (2008) A Digital Dictionary of Catalan Derivational Affixes. In proceedings of Euralex 2008, Barcelone, Spain, IULA

Cartoni B. and Lefer M.A. (2009), Improving the Representation of Word-Formation in Multilingual Lexicographic Tools: the MuLeXFoR Database, In proceedings of Euralex 2010, Leewaerden, Netherland.

Cartoni B. (2009), Lexical Morphology in Machine Translation: a Feasibility Study, In proceedings of EACL 2009, Athens, pp. 130-138.
Cartoni, B. (2008a) La place de la morphologie constructionnelle dans les dictionnaires bilingues: étude de cas. In proceedings of Euralex 2008, Barcelone, Spain, IULA: pp. 813-820.
Cartoni B. (2008b) De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue. University of Geneva: Geneva. PhD Thesis.
Dardano, M., Frenguelli G. and Colella G. (2006) What Lexicographers Do with Word Formation. In proceedings of Euralex 2006, Torino: pp. 1115-1127.
Fradin, B. (2003) Nouvelles approches en morphologie. Paris, Puf.
Iacobini C. (2004) I prefissi. In Grossmann M. and Rainer F. (eds) La formazione delle parole in italiano. Niemeyer: Tübingen, 99-163.
Lefer, M.-A. (2009) Exploring lexical morphology across languages: a corpus-based study of prefixation in English and French writing. Louvain-la-Neuve, Université catholique de Louvain. PhD Thesis
Namer, F. (2009) Morphologie, lexique et traitement automatique des langues : l'analyseur DériF. Paris, Lavoisier.
Prcic, T. (1999) The treatment of affixes in the "big four" EFL dictionaries. International Journal of Lexicography 12(4) pp. 263-279.
Szymanek B. (1988) Categories and Categorization in Morphology. RW-KUL: Lublin, PhD Thesis
ten Hacken, P., Abel A. and Knapp J. (2006) Word Formation in an Electronic Learners' Dictionary',. International Journal of Lexicography.

Appendix

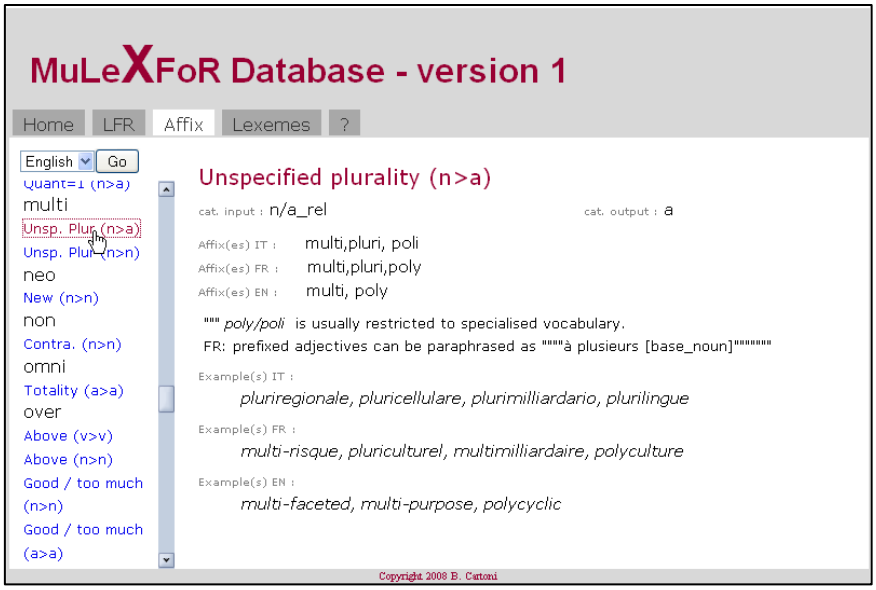

Figure 2: Browsing by affix

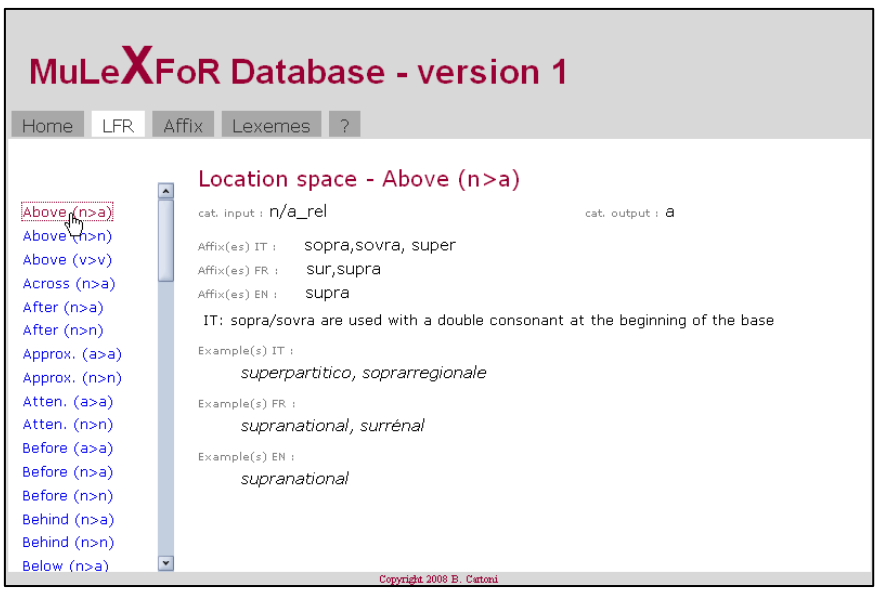

Figure 3: Browsing by rule