MORBO/COMP: a multilingual database of compound words

Emiliano Guevara[*], **Sergio Scalise**[*], **Antonietta Bisetto**[*], **Chiara Melloni**[†]

[*]Università degli studi di Bologna
Dip. di lingue e letterature straniere moderne

[†]Università degli studi di Verona
Dip. di germanistica e slavistica

Contact: Emiliano Guevara, emiliano@lingue.unibo.it
Via Cartoleria n. 5, 40124 Bologna (Italy)

Abstract

The aim of this paper is to present the MORBO/COMP project, which has reached its final stage in development and will soon be published on-line. MORBO/COMP is large database of compound types in over 20 languages. The data for these languages have been collected and analysed by a group of morphologists from various European countries.

1. Introduction

Scientific research on word formation at the international level has dedicated a great deal of attention to compounding phenomena. The literature knows of several detailed studies on the subject, in particular: on English (among others) cf. Marchand (1969), Selkirk (1982), Bauer (1983); on Italian cf. Scalise (1992, 1994), Bisetto (2004); on Spanish cf. Bosque & Demonte (1999); French, Villoing (2002), Di Sciullo (1992), Zwanenburg (1970); German, cf. Becker (1992); Chinese, cf. Ceccagno & Scalise (2005), etc.

These works have contributed to increase our knowledge of the various existing types of compounds, although from a language-specific point of view. For instance, data on the degree of endocentricity/exocentricity in the world's languages is not available yet. There is no reliable source of data describing the different attested types of compounds, the structural complexity of possible compound words, the presence and typology of linking elements, plural formation, distribution of different structures in the world's languages, whether categorial and semantic head coincide, etc. Therefore, general studies of compounding, classification of compound types, of the typology of compounding in the world's languages have not been undertaken.

MORBO/COMP seeks to establish itself as the first empirical base on the basis of which such studies on compounding will be possible.

2. Compounding the problem

A systematic compilation of compounding data allowing interlinguistic comparison does not exist. As a result, every hypothesis proposed so far is descriptively inadequate and language-specific. For instance, data on the degree of endocentricity/exocentricity in the world's languages is not available yet. There is no reliable source of data describing the different attested types of compounds, the structural complexity of possible compound words, the presence and typology of linking elements, plural formation, distribution of different structures in the world's languages, whether categorial and semantic head coincide, etc.

The preliminary problem that arises is the delimitation of the field of investigation. It is not clear, at present, which are the boundaries of compounding phenomena. A starting definition is provided by Bauer (2001, 695):

> "a lexical unit made up of two or more elements, each of which can function as a lexeme independent of the other(s) in other contexts, and which shows some phonological and/or grammatical isolation from normal syntactic usage."

This definition is, however, problematic because it cannot always discriminate between compounds and phrases. It is also possible to apply additional criteria (e.g. syntactic atomicity, semantic specialization, etc.), though never reaching definitive results.

After the pioneering work of Ten Hacken (1994), research has been limited to the mere analysis of single compound types (e.g. Olsen, 2001), not being able to reach significant levels of descriptive adequacy, especially from a typological point of view. Thus far, therefore, research on compounding phenomena could not count on neither a consistent theoretical framework nor a sound empirical base beyond the language under consideration.

With the exception of Bauer (2001), research on the typology of compounding has been extremely meagre, if not totally neglected. Moreover, while a good number of international research projects on syntactic typology have emerged (such as Eurotyp – University of Konstanz – and The World Atlas of Language Structures – WALS, Max Planck Institute, Leipzig –), there is currently not a single initiative on the typology of compounding. Linguistic typology has always paid scarce attention to non-inflectional morphological processes.

Furthermore, typological frameworks built on correlations between compounding strategies and other typologically relevant parameters (i.e. word order, head-position, morphological type) are very rare (an exception is Bauer, 2001).

Another crucial aspect of the theoretical treatment of compounds regards their classification: the criteria traditionally used to classify compounds are flawed in the following fundamental aspects:

i. they are usually language-specific, and

ii. they are not internally coherent.

Regarding (i), Anglo-Saxon classifications (root / primary compounds vs. verbal / secondary compounds, cf. Marchand, 1969; Allen, 1978; Selkirk, 1982) are valid for Germanic languages, but not for Romance languages, for instance. Also, the traditional Sanskrit classification of compounds have been applied to European languages without success. For example, the term 'dvandva compound' has been incorrectly used for Germanic and Romance languages (the real Sanskrit dvandva compounds do not have a unique semantic referent, unlike the Germanic and Romance words analyzed as such).

With respect to (ii), very heterogeneous criteria have been used to classify compounds (Marchand, 1969; Olsen, 2001; Bauer, 2002; Haspelmath, 2002; Booij, 2005). In particular, the syntactic notions of subordination and coordination (which regard the relation between the two constituents of the compound) are traditionally considered to be at the same level as notions such as endocentric / exocentric (which, instead, describe the relation between the head of a compound and the compound as a whole). Clearly, these notions belong to different levels of analysis, since coordinate and subordinate compounds can be both endocentric and exocentric.

The MORBO/COMP project started as a first initiative to answer to these problems. Its aim is to collect compounding data in a standardized manner, to allow cross-linguistic comparison. The database is meant to serve the needs of researchers working on the theoretical and typological aspects of compounding. However, we are interested in further developing the database as a knowledge source for data-driven research: MORBO/COMP, providing a concise and detailed description of the possible compounds in over twenty languages, could be the basis for automated tasks such as compound recognition and classification in large corpora. Similarly, the database could be used as a source for example-based activities in various kinds of experimental research.

Naturally, the database cannot possibly contain answers to problems of a theoretical nature, but it will certainly help researchers look for the solutions, providing them with a sound empirical base.

3. Description of the database

The MORBO/COMP database is organized in three distinct levels. For every studied language there is:

i. an EXPANSION listing all the collected compounds,

ii. a TABLE where four examples of each compound type are represented and analyzed in 20 different fields (comprising categorial, structural, morphosyntactic and semantic information),

iii. a SKELETON, summarizing the possible structures of a given language.

Each of the fields describing a compound in the TABLE data-set can be used as a search-key to query the database, interlinguistically and intralinguistically.

The relational architecture of the database permits a broad range of immediate findings, e.g. whether a language or a group of languages has or does not have compounds belonging to a particular lexical category; whether a language has or does not have a particular structural type (e.g. [V+N] is present both in Chinese and in Italian); which is the canonical position of the head constituent in each of the different languages.

Similarly, the database can give immediate answers to many other questions, such as: are there languages with [V+N] constructions with an eventive reading? Which are the possible combinations of the lexical categories for a given language? Which are the preferred strategies for morphosyntactic marking of compounds for a given language, etc. These are only a few of the many possibilities.

3.1. Database fields

Compounds in the MORBO/COMP database (TABLE subset) are analyzed in the following fields:

- Language
- Compound
- Category: N, V, A, P, Adv, etc.
- Structural description: e.g. [N+A], [V+N], etc.
- Classification: 3 major classes: SUB (subordinative, e.g. *truck driver*), CRD (coordinative, e.g. *king emperor*), ATAP (appositive / attributive, e.g. *key word*).
- Endocentricity
- Categorial / syntactic head: right left, both, none (respectively, *high school*, it. *capo stazione* 'station master', *deaf-mute*, *pickpocket*.
- Semantic head: idem
- Constituents
- Category of constituents: N, V, A, P, Adv, etc.
- Linking elements
- Morphosyntactic marking: 1 (on the first constituent), 2 (on the second constituent), 12 (on both constituents), 0 (invariable)
- Gender: gender of the first contituent (m., f.., n., m./f., 0) + gender of the second constituent (id.) = gender of the whole (id.). E.g. capostazione m+f=m.
- Gloss: gloss of the compound and of the constituents in English

3.2. Languages included in the database

At present, MORBO/COMP includes compounding data for the following languages:

Basque
Byelorussian
Chinese
English
French
Greek
Hebrew
Italian
Korean
Norwegian
Portuguese
Serbo-croatian
Swedish
Bulgarian
Catalan
Dutch
Finnish
German
Greek
Hungarian
Japanese
Latin
Polish
Russian
Spanish
Turkish

Each language is represented by data-sets that differ in granularity and coverage. It is expected that the final stage of development of the electronic resources will vary for each language, but the eventual negative side-effects will be minimized thanks to a flexible database system and interface able to ensure optimal use also of partial resources.

4. Dissemination

The MORBO/COMP database, once completed, will be published online. The entire amount of data acquired by the project's activities will be made available to the general public with an intuitive and flexible web-interface.

Currently, the database interface to MORBO/COMP is under development. It has been projected with a net-like structure, allowing retrieval of data on various levels of complexity. For example, it will be possible to query the database by language and/or type of compound (internal structure, constituent part of speech, classification, etc.), but also by specific (parts of) compound words.

5. Typological coverage

As it has been pointed out above, MORBO/COMP aims at becoming a useful instrument for the typological study of compounding. However, our group has experienced great difficulties in obtaining enough data to achieve an adequate description of compounding phenomena in the currently represented languages.

Traditionally, typological surveys are based on written sources: dictionaries and grammars. In this way, a high number of languages, well-balanced from the typological and areal point of view, is relatively easy to achieve.

This methodology proved to be useless to collect compounding data: traditional written sources usually do not include enough examples of the various structural patterns and/or classes. The MORBO/COMP database has relied heavily on native speakers' work to collect, classify and analyze manually all the represented examples. Unfortunately, this approach turns to be quite slow and costly. At present, MORBO/COMP is not typologically well-balanced. We are well aware of this shortcoming, and are currently trying to extend the database coverage to extra-european languages, with the final aim of reaching a typologically significant language sample.

6. Concluding remarks

Morphological theory will surely benefit from such a specialized database. The richness of its data could serve as basis for new theoretical representations and classifications of compounds. Only a brief account of the database's general features has been provided here. In the next months of the project will be published online.

Although the project has reached a final stage in the development, the framework will remain open for future additions and enhancements, especially with the aim of improving the typological coverage of the database.

7. References

Bauer, L. (2001), Compounding. In: M. Haspelmath et al. (eds.) *Language Typology and Language Universals*. An International Handbook. Berlin - New York. Walter de Gruyter. 695-707.

Becker, Th. (1992), Compounding in German. Rivista di Linguistica 4/1, 5-36.

Bisetto, A. (2004), Composizione con elementi italiani. In: M. Grossmann & F. Rainer (eds.), *La formazione delle parole in italiano*, Tübingen, Niemeyer, 33-51, 53-55.

Bisetto, A. & S. Scalise. (2005). The classification of compounds. *Lingue e Linguaggio*, IV.2. 319-332.

Bosque, I. & V. Demonte (eds.)(1999) *Gramática descriptiva de la lengua española*. Madrid, Espasa Calpe.

Ceccagno, A. & S. Scalise (2005), Composti del cinese: analisi delle strutture e identificazione della testa. In: A.M. Palermo (ed.) *La Cina e l'altro*, Napoli, Il Torcoliere, 1-29.

Di Sciullo, A.M. (1992), Deverbal compounds and the external argument. In: I.M. Roca (ed.), *Thematic Structure: its Role in Grammar*, Berlin-New York, Foris, 65-78.

Marchand, H. (1969). *The Categories and Types of Present-day English Word-formation*. München: C. H. Beck.

Olsen, S. (2000), Composition. In: G. Booij et al. (eds.), *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung*, Berlin, de Gruyter, 897-916.

Olsen, S. (2001), Copulative Compounds. A closer look at the interface between morphology and syntax. In: G. Booij & J. van Marle (eds.) *Yearbook of Morphology* 2000, 279-320.

Packard, J. (2000). *The Morphology of Chinese*, Cambridge: Cambridge University Press.

Ralli, A. (1992), Compounding in Greek. Rivista di Linguistica 4/1, 75-98.

Scalise, S. (1992) (ed.), Special Issue of *Rivista di Linguistica* on Compounding, Turin, Italy, 4 (1).

Scalise, S. (1992), Compounding in Italian. Rivista di Linguistica 4/1, 175-198.

Scalise, S. (1994), *Morfologia*, Bologna: Il Mulino.

Selkirk, E. (1982). *The Syntax of Words*. Cambridge (Mass.): MIT Press.

Villoing F. (2002), *Les mots composeés [VN] du français*. Ph.D. dissertation, Paris Nanterre.

Zwanenburg, W. (1992), Compounding in French. Rivista di Linguistica 4/1, 221-240.