INFLECTIONAL NON COMPOSITIONALITY AND VARIATION OF COMPOUNDS IN FRENCH, POLISH AND SERBIAN, AND THEIR AUTOMATIC PROCESSING

Agata Savary
Laboratoire d'Informatique
François-Rabelais University of Tours, France

Cvetana Krstev, Duško Vitas
Faculty of Philology and Faculty of Mathematics
University of Belgrade, Serbia

Abstract

Compounds and other multi-word units are of essential qualitative and quantitative importance in natural languages. We address their linguistic properties in French, Polish and Serbian with respect to their inflectional morphology. We are particularly interested in the cases of morphological non compositionality. We also account for some orthographic, syntactic and semantic variants of compounds. We show how most of the linguistic properties studied can be formally described within the framework of a lexicalized formalism, such as the *Multiflex* system, meant for an automatic treatment of the computational morphology of compounds.

Key-words

Compounds; multi-word units; inflectional morphology; computational morphology; variation; Slavic morphology

Résumé

Les mots composés et autres unités polylexicales sont d'une importance qualitative et quantitative capitale dans les langues naturelles. Nous nous intéressons aux propriétés linguistiques liées à la flexion, en français, polonais et serbe, et plus particulièrement au phénomène de non compositionnalité morphologique. Nous étudions également la variation orthographique, syntaxique et sémantique. Tout au long de cette étude, nous montrons comment la plupart des propriétés étudiées peuvent être

formellement décrites par un formalisme lexicalisé implémenté dans le système *Multiflex* dédié au traitement automatique de la morphologie computationnelle.

Mots clés

Mots composés ; unités polylexicales ; morphologie flexionnelle ; morphologie computationnelle ; variation, morphologie slave

1. Introduction

There is a growing interest, within the community of computational linguists, in the automatic processing of compounds, and other multi-word units, because of their essential qualitative and quantitative importance in natural languages.

Similarly to simple words, compounds are subject to inflection. Obviously, a reliable description of the inflection of single words is a necessary condition for the description of the inflection of compounds. However, this condition is rarely a sufficient one. For example, in order to obtain the plural form of the English compounds:

(1) *chief justice*

(2) *lord justice*

not only do we need to be able to generate the plural of *chief*, *lord* and *justice* but we also need to know how the different inflected forms of these constituents combine. For instance the plural forms are correct in:

(3) *chief justices*

(4) *lord justices*, *lords justice*, *lords justices*

but not in:

(5) **chiefs justice*, **chiefs justices*

In English, such examples belong to a closed list and are of relatively little quantitative importance. Since French presents a richer inflectional morphology, inflectional irregularities within compounds are frequent. For instance, the class of French *Verb-Noun*-type compounds contains numerous examples in which the gender and number of the whole structure cannot be deduced from those of its constituents. For instance the French compound:

(6) *un perce-neige* (‘a snowdrop’)

is masculine although the noun *neige* is feminine.

In Slavic languages, the difficulties with the inflection of compounds may be even more important due to several factors: (i) declension and a complex gender, number and animateness cross-references within nouns and adjectives, (ii) the size of the inflectional paradigm, and (iii) a relatively free word order in nominal phrases among others.

In [Goussier, to appear] we have performed a detailed study of the linguistic properties of compounds with respect to inflection in several European languages. We have also proposed [Savary 2005a] a formalism for a *lexicalized* (compound-per-compound) description of inflectional morphology, as well as the accompanying *Multiflex* system which allows for an automatic generation of inflected forms of compounds.

In the present paper we introduce the *Multiflex’* formalism, then we review some linguistic properties of compounds, with a particular impact on French, Polish and Serbian. We recall the notion of headword and of morphological compositionality, which calls for an efficient representation of agreement rules via unification. Then we study in detail the morphological non compositionality of compounds, which is one of their defining criteria. We also consider the problem of graphical, inflectional, syntactic and semantic variations of compounds, which should possibly be considered in the same framework as their inflected forms. We illustrate our considerations with numerous examples, each of which is labelled with the abbreviation of the language it comes from: French (FR), Polish (PL), Serbian (SR). The formal description of some examples is given within the *Multiflex* formalism.

2. Compounds – scope of the study

Compounds encompass a bunch of hard-to-define and controversial linguistic objects ([Habert and Jacquemin 1993], [Corbin 1992]). Their numerous linguistic and pragmatic definitions ([Benveniste 1974], [Downing 1977], [Levi 1978], [Bauer 1983], [Gross 1990], [Anscombre 1990], [Silberztein 1993a], [Gross 1996], [Cadiot 1992]) invoke three major points:

- they are composed of two or more words

- they show some degree of morphological, distributional or semantic non-compositionality
- they have unique and constant references

However, the basic notions (a word, a reference, non-compositionality) and measures (degree of non-compositionality) used in those definitions are themselves controversial.

In this paper, we define the scope of our work pragmatically: we consider a compound as a *contiguous* sequence of graphical units which, for some application-dependent reasons, has to be listed, described (morphologically, syntactically, semantically, etc.) and processed as a unit. Such sequences usually have the forms of nominal, adjectival, prepositional or adverbial phrases. The first two are of particular interest here since they are concerned by inflection.

This large application-dependent view allows one to address the in-vogue notion of *multi-word units*, used in computational linguistics with respect not only to traditional compounds but also to other linguistically close objects, such as complex terms and named entities.

Furthermore, we limit ourselves to the formal description of the *inflectional paradigms* of compounds. For the purpose of their automatic treatment we are interested in their fully reliable description, i.e. for each compound, all the correct inflected forms should be described and no incorrect form should be described and no incorrect form should be allowed.

Regular inflection cases may be accounted for by general grammar rules such as: “In French, in a *Noun-Preposition-Noun* structure, the first noun is the headword. In order to inflect the compound one needs to inflect the headword and leave all other components unchanged”. As soon as a compound may be identified as having the corresponding structure, the generation of its plural is straightforward, e.g.:

(7) (FR) *toile de fond > toile<u>s</u> de fond* (‘backdrop in a theatre’)

However, the inflection of some compounds, even having the same headword, may be idiosyncratic, as in:

(8) (FR) *toile d'araignée - toile<u>s</u> d'araignée, toile<u>s</u> d'araignée<u>s</u>* ('a spider's web')

These two examples show that the inflectional difference between compounds sharing the same structure is a phenomenon which occurs at the lexical level rather than at the grammatical level. The difference between *toile de fond* and *toile d'araignée* may only be expressed as an exceptional behavior of *araignée* in relation to *toile*.

In this paper we are particularly interested in *inflectionally non compositional compounds*, such as (6) and (7). Their fully correct description calls for a lexicalized approach, such as the *Multiflex* formalism, in which each compound, whether compositional or non compositional, obtains an explicit inflection pattern which determines all of its correct inflected forms. We perform a linguistic study of the conditions in which inflectional non compositionality occurs in French, Polish and Serbian, and show how they can be accounted for in *Multiflex*.

3. Describing inflectional properties of compounds in *Multiflex*

As mentioned in [Savary 2005a] and [Goussier, to appear], *Multiflex* is an implementation of a formalism meant for an exhaustive, correct, and compact description of the inflectional behavior of compounds. It is based on the abstract morphological representation of simple words, as well as some regular-language patterns and unification agreement rules.

It is relatively independent of the morphological model which applies to simple words, in the sense that it may work with any underlying module handling the inflection of simple constituents, as long as this module respects some interface constraints.

3.1. Graphical aspects

The main constraint for such a system is to be able to define the notion of graphical unit formally, and to inflect the units that it has itself defined. For instance it might identify the compound:

(9) (FR) *bonhomme de neige* ('snowman')

as a sequence of 5 graphical units (*bonhomme*, blank space, *de*, blank space, *neige*) or as 6 graphical units (*bon*, *homme*, blank space, *de*, blank space, *neige*), on the condition that it allows to generate the plurals, *bonshommes*, or *bons* and *hommes*, respectively.

Within *Multiflex*, the corresponding graphical units of a compound lemma are referred to via ordered variables *$1*, *$2*, *$3*, etc. Thus, compound (9) may be represented graphically as in Fig. 1 or Fig. 2, depending on the underlying module used to describe the morphology of simple words.

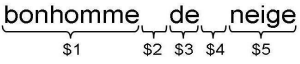

Fig. 1. One possible graphical division of *bonhomme de neige*

Fig. 2. Another possible graphical division of *bonhomme de neige*

In most of the examples presented in the following sections, *Multiflex* uses the Unitex' [Paumier 2002] convention, in which a graphical unit is either a maximum-length contiguous sequence of alphabet characters, or a single non alphabet character. Thus, example (9) is decomposed as in Fig. 1.

3.2. Inflectional paradigm of a compound in Multiflex

The inflectional paradigm of a compound is the list of its inflected forms together with their inflectional description. As explained in [Goussier, to appear], the size and contents of this list depend not only on the nature of the language studied, but also on the morphological model chosen for the given language.

That aspect is particularly visible in Slavic languages. In Polish, the gender category is seen as a "flat" list of features, or as a hierarchy [Przepiórkowski 2003]. The number of genders is estimated at five [Przepiórkowski 2004], six [Szpakowicz 1986 and Vetulani et al. 1998], eight [Woliński 2001] or eleven [Jassem 1996]. In some of these models inflection is a matter of the Cartesian product of all relevant features, while in [Vitas, Krstev 2001] for Serbian, only some combinations of number, gender, case and animateness values are allowed. Agreement rules are seen as straightforward equations of features, or involve specialized algorithms such as no-care symbols [Vitas, Krstev 2001] or complex multi-tuple relations [Jassem 1996].

Whatever is the morphological model chosen for Polish or Serbian, the corresponding inflection paradigms of nouns and adjectives are always rather huge, as they count several dozens of forms, or even over a hundred of them. Thus, they call for mechanisms allowing a compact representation, such as *unification* and *value propagation* discussed below.

In *Multiflex*, the French, Polish and Serbian inflection categories and values may be represented by the textual descriptions whose fragments are shown in Fig. 3, Fig. 4, and Fig. 5, respectively. The labels used for categories and values may be freely chosen, nonempty sequences of characters.

```
Nb: sing, pl
Gen: masc, fem
```

Fig. 3. Some inflectional categories and values for French

```
Nb: sing, pl
Gen: masc_pers, masc_anim, masc_inanim, fem, neu
Case: Nom, Gen, Dat, Acc, Inst, Loc, Voc
```

Fig. 4. Some inflectional categories and values for Polish

```
Nb: s, p, w
Gen: m, f, n
Case: 1, 2, 3, 4, 5, 6, 7
Anim: v, q, g
Comp: a, b, c
Det: d, k, e
```

Fig. 5. Some inflectional categories and values for Serbian

The inflection paradigm of a compound is represented in *Multiflex* via an *inflection graph*. Let us consider the following example:

(10) (FR) *mémoire vive à progresseurs* (‘ self-incremental random access memory’)

and the corresponding inflection graph in Fig. 6. Each node in the graph describes an action to be performed on a single constituent, numbered *$1, $2, $3*, etc. (cf section 3.1). The action may correspond either to recopying the constituent with no change, or to generating its particular inflected form. For instance, *<$4>* means recopying the fourth constituent (here: the blank space), *<$5>* means recopying the fifth one (here: *à*), etc., while *<$3:Nb=pl>* means that the third constituent (here: *vive*) should be inflected in the plural, while keeping the same gender (here: feminine). The category and the value identifiers must be the same as in the language-level description file (Fig. 3).

Each path in a graph starts with the leftmost arrow and ends with the rightmost encircled box. While following a path, we perform the action contained in each box, and we accumulate the morphological features indicated below the boxes (here: <*Gen=fem;Nb=sing*> or <*Gen=fem;Nb=pl*>). The result of the accumulated features gives the complete morphological description of the inflected forms of a compound. Thus, the application of the graph in Fig. 6 to compound (10) results in the following inflection paradigm:

(11) (FR) *mémoire vive à progresseurs:* <Gen=fem;Nb=sing>
(FR) mémoires vives à progresseurs: <Gen=fem;Nb=pl>

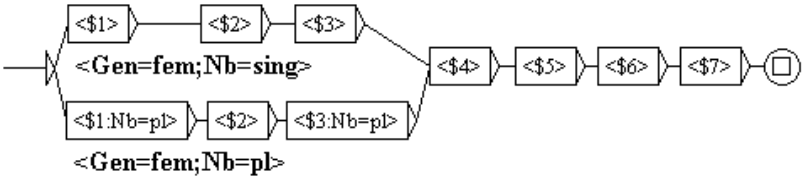

Fig. 6. Reduced inflection graph for *mémoire vive à progresseurs* in French

Agreement rules may be expressed in *Multiflex*' graphs via explicit value assignments, such as *Nb=pl* for *$1* and *$3* in Fig. 6, or via **unification**. In Fig. 7, the two paths from the previous graph have been replaced by one path, in which the number inflection is expressed via the unification variable *$n*. A unification variable, i.e. an alphanumerical identifier preceded by the dollar sign, may take any value from the domain corresponding to the inflection category it is assigned to: here, *$n* may take any value of *Nb* among those listed in Fig. 3, i.e. *sing* or *pl*. The instantiation of the variable is unique in the whole path: if we fix *$n* to *sing* for the first constituent, then the same value has to be set for the third one, as well as for the whole compound. Similarly, if we fix *$n* to *pl* while processing the first node, it has to remain *pl* until the end of the path.

The inflected forms resulting from the exploration of this graph are identical to (11). However the graph in Fig. 7 is more general than the one in Fig. 6 in that it can be applied both to (10) and to (12):

(12) (FR) *pont suspendu à haubans* ('cable-stayed bridge')

due to the **value propagation** mechanism expressed by a double equal sign (==). If a unification variable is assigned to a category with this sign then it inherits the value of this category from the corresponding constituent, as it appears in the base form of the compound. Thus, the first box in Fig. 7 fixes variable *$g* to the feminine for (10), and to the masculine for (12), because *mémoire* and *pont* are feminine and masculine, respectively.

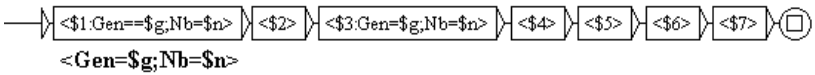

Fig. 7. Inflection graph with unification and value propagation for *mémoire vive à progresseurs* and *pont suspendu à haubans* in French

Note that Fig. 7 clearly shows which constituent is the headword: it is the one in which the value propagation takes place. The common unification variables *$g* and *$n* express the agreement rules between the headword and the characteristic constituent (*$g* in box *$3* is optional here because gender never changes). Finally, the occurrence of the same variables below the path shows that the gender and the number of the whole compound are determined by those of the headword.

The possibility of expressing agreement rules via a unification mechanism is crucial in Slavic languages. If this mechanism were not available, each inflected form of a compound would have to be represented by a separate path, as in Fig. 6 for French. As a result, the inflection graph would have to contain dozens of distinct paths, which would make it unreadable and unmaintainable. With unification, most regular compounds can be described by a single path in Polish, and by less than 5 paths in Serbian (due to particularities concerning the no-care values). For instance the compound adjectives:

(13) (PL) *jaki taki* ('passable, tolerable')

(14) (SR) *sam samcit ('completely alone')*

require graphs containing a unique path, in which all inflection values are unified via common variables. They would have to contain 70 and 77 paths, respectively, if no unification were allowed.

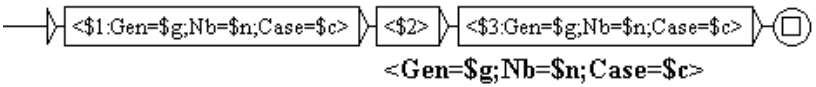

Fig. 8. One-path graph describing 70 inflected forms of *jaki taki* in Polish

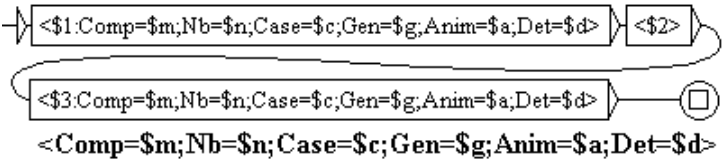

Fig. 9. One-path graph describing 77 inflected forms of *sam samcit* in Serbian

4. Morpho-syntactic non compositionality

Compounds (10) through (14) are inflectionally compositional, in the sense that their inflectional properties can be fully deduced from the properties of their respective constituents. In each form the compound has the same inflection features as its headword, and the "usual" agreements take place between the headword and the so-called *characteristic constituents* [Silberztein 1990], here *vive* and *suspendu*. The non characteristic constituents (*progresseurs, haubans*) never vary.

However, it is a well known fact that in most cases compounds are, at least partly, non compositional from a morphological, syntactic, semantic, or distributional point of view. This fact is considered as a *defining criterion* of compounds with respect to free structures by [Gross 1988].

Inflectional non compositionality may be provoked by any of the facts detailed in sections 4.1 through 4.4.

4.1. The phrase has no headword

Such a phrase is a so-called *exocentric* phrase, i.e. it contains no word from which its inflectional properties might be deduced.

In French such cases are frequent in *Verb-Noun* structures, such as:

(15) (FR) *un porte-serviettes* ('towel rack')

as well as in nominal structures in which the head nouns have disappeared, as in:

(16) (FR) *une deux chevaux* (literally : 'a two horses'; initially *une voiture à deux chevaux* 'a car with a two-horsepower engine')

Note that the nouns *serviettes* and *chevaux* cannot be head nouns since they are in the plural, feminine and masculine, respectively, while the compounds are in the singular, masculine and feminine, respectively. Since the compounds are identical in both numbers, the corresponding graphs (Fig. 10 and Fig. 11) contain only one unification variable *$n*, which agrees with no constituent.

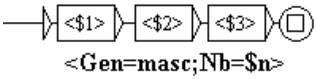

Fig. 10. Inflection graph for an exocentric masculine compound such as *porte-serviettes* in French.

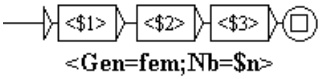

Fig. 11. Inflection graph for an exocentric feminine compound such as *deux chevaux* in French.

Note also that due to the absence of a headword, the gender value for these compounds must be indicated explicitly, as it cannot be inherited from any constituent. Thus, the two graphs cannot be merged into one by replacing the gender values by a unification variable, as was the case with the graph in Fig. 7. If we did merge the graphs, *porte-serviettes* and *deux chevaux* would erroneously be allowed to inflect in gender (**une porte-serviettes*, **un deux chevaux*).

In Slavic languages headless compounds are less frequent but do exist:

(17) (PL) *sam na sam* (literally 'alone-against-alone')

Example (17) functions as a noun in neuter gender, although it contains no neuter noun. It might result from the underlying "free" structure:

(18) (PL) *spotkanie sam na sam* ('a meeting in private between two individuals')

All the inflected forms for this compound are identical in all numbers and cases, and can be described by the one-path graph in Fig. 12.

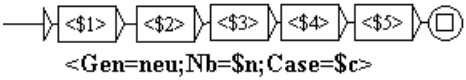

Fig. 12. Inflection graph for a headless neuter compound such as *sam na sam* in Polish.

4.2. The "usual" number, gender, case, etc. agreements are not always fulfilled

In some cases the constituents which are supposed to be characteristic do not always agree with the headword.

For instance, the compound:

(19) (FR) *une grand-mère, des grand-mères, des grands-mères* ('grandmother')

is of *Adjective-Noun* structure in which the adjective typically agrees with the noun in gender and number. However, number agreement is fulfilled here only in one of the two plural variants (*grands-mères*), and gender agreement does not take place at all: *grand* is masculine, although the noun it modifies is feminine. Thus, the upper path in the corresponding graph in Fig. 13 accounts for the forms in which the adjective does not agree with the head noun (*grand-mère, grand-mères*), while the lower path describes the partial (number-only) agreement between these two constituents in the plural (*grands-mères*).

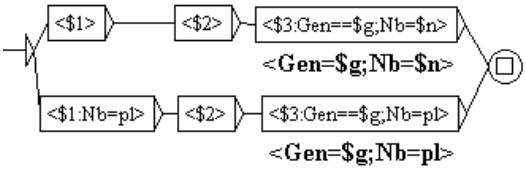

Fig. 13. Inflection graph accounting for an irregular partial agreement between the head noun and its modifier as in *grand-mère* in French.

4.3. At least one inflected form that is usually expected for the structure under consideration is nonexistent.

For instance, the compound:

(20) (PL) *zimne nogi* (literally 'cold legs'='a dish consisting of meat and jelly')

does not admit a singular form, even if *zimna noga* is a syntactically correct form (in the singular the phrase loses its terminological sense). Thus, in Fig. 14, the gender and number values are fixed and inherited from the head noun (here: *nogi*), while the case varies.

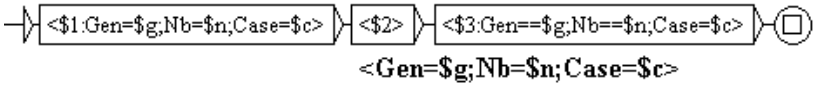

Fig. 14. Accounting for the nonexistence of the number inflection in compounds like *zimne nogi* in Polish.

Note that the nonexistence of a particular inflected form is not always a proof of the inflectional non compositionality of a compound, as it may simply result from the inflection restrictions of the headword. For instance, the compounds:

(21) (FR) *funérailles nationales* ('national funeral')

(22) (PL) *krótkie spodnie* (literally 'short trousers' = 'shorts')

(23) (SR) *crne naočari* (literally 'black glasses'='sunglasses')

do not admit a singular form due to the fact that their head nouns *funérailles*, *spodnie* and *naočari* are plural-only nouns. Thus, example (22) can be described equally by Fig. 14 or by Fig. 15, in which the number value is not fixed for *$3*. If *$n* gets instantiated with the singular, the corresponding singular compound form is simply omitted because the singular form of its headword does not exist.

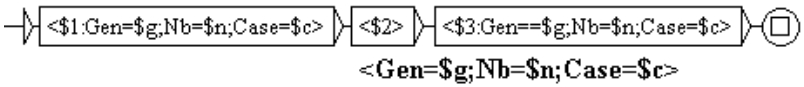

Fig. 15. Accounting for regular inflection in compounds containing plural-only head nouns like *krótkie spodnie* in Polish.

4.4. At least one inflected form is irregular

A form is irregular if we need to inflect a constituent that does not normally agree with the headword (i.e. a non characteristic constituent).

Appositions, i.e. particular types of *Noun-Noun* phrases, are frequent examples of such irregularities. For instance, the following compounds:

(24) (FR) *bateau-mouche* (literally 'a fly boat' = 'a Paris-style river boat')

(25) (PL) *człowiek kot* (literally 'cat man'='a very agile person')

(26) (SR) *čovek žaba* (literally 'man frog'='a man specially dressed and equipped to explore the bottom of the sea')

possess head nouns in initial position, whose inflectional features they inherit: masculine singular, nominative masculine human singular, and nominative singular masculine animate, respectively. The second noun in each example has a different gender than the head noun: *mouche* is feminine, *kot* is animate masculine, and *žaba* is feminine. However, these constituents do agree in number or case with the head:

(27) (FR) *bateau<u>x</u>-mouche<u>s</u>*

(28) (PL) *człowiek<u>a</u> kot<u>a</u>, człowiek<u>owi</u> kot<u>u</u>, etc.*

(29) (SR) *čovek<u>a</u> žab<u>e</u>, čovek<u>u</u> žab<u>i</u>, čovek<u>a</u> žab<u>u</u>, etc.*

Thus, in the graph in Fig. 16, which describes example (25), the gender of the third constituent (here: *kot*) is not unified with the gender of the first one, while the case and number agreements are expressed.

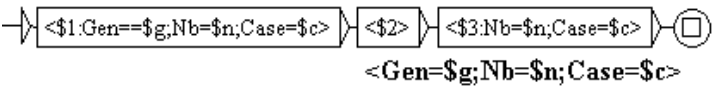

Fig. 16. Accounting for the inflection of a non characteristic constituent in compounds like *człowiek kot* in Polish.

A few compounds admit two variants of an inflected form, one of which is regular, and the other one which involves the inflection of a non characteristic constituent (here: *araignée*, *Ostoja* or *tetke*):

(30) (FR) *toile d'araignée* ('spider's web', singular)
**toile d'araignées* (false singular)
toiles d'araignée, toiles d'araignées (plural)

(31) (PL) *Ostoja-Malinowskiego, Ostoji-Malinowskiego* (family name in the singular)
Ostoja-Malinowscy (plural)
**Ostoje-Malinowscy* (false plural)

(32) (SR) *sestra od tetke* (literally 'sister from aunt' = 'daughter of one's mother's or one's father's sister', singular)
**sestra od tetaka* (false singular)
sestre od tetke, sestre od taka (plural)

Thus, example (30) is described by two paths in Fig. 17: one in which the last constituent (here: *araignée*) remains the same, and the other in which it agrees in number with the head noun (here: *toile*).

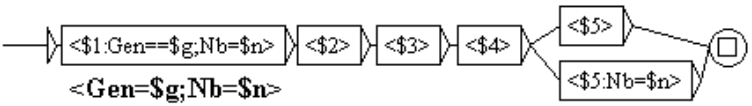

Fig. 17. Inflection graph accounting for two variants of the plural in *toile d'araignée* in French.

In some cases each inflected form may count two variants (in which the non head word is either singular or plural), as in:

(33) (FR) *simulateur de vol, simulateur de vols* ('flight(s) simulator', singular)
simulateurs de vol, simulateurs de vols (plural)

(34) (PL) *pranie mózgu, pranie mózgów* ('brainwashing', singular)
prania mózgu, prania mózgów (plural)

(35) (SR) *nosilac spomenice, nosilac spomenica* (literally 'carrier of distinction(s)' = 'distinguished war veteran', singular)
nosioci spomenice, nosioci spomenica (plural)

Thus, in Fig. 18 two different unification variables *$n1* and *$n2* appear for the first and the third constituent. These variables may be instantiated with *sing* or *pl* independently of each other, which results in four inflected forms

for each case. The first constituent being the headword, it provides the gender, number and case values to the whole compound.

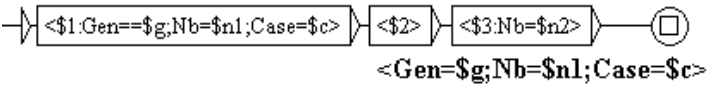

Fig. 18. Accounting for an independent number inflection between the head noun and the modifier as in *pranie mózgu* in Polish.

Similar examples may be found concerning compound adjectives. For instance, the two equivalent Polish and Serbian compounds:

(36) (PL) *głodny jak wilk* ('as hungry as a wolf')

(37) (SR) *gladan kao vuk* ('as hungry as a wolf')

possess two variants of the plural (and paukal in Serbian), in which the nouns *wilk* and *vuk* may or may not agree with the head adjective in number:

(38) (PL) *głodn<u>i</u> jak wilk<u>i</u>, głodn<u>i</u> jak wilk*

(39) (SR) *gladn<u>i</u> kao vuk<u>ovi</u>, gladn<u>i</u> kao vuk*

The upper paths in the graphs in Fig. 19 and Fig. 20 describe the singular forms. Thus, the fifth constituent is marked as invariable. The other paths, corresponding to the plural (*pl* for Polish, *p* for Serbian) and to paukal (*w* for Serbian), admit a number inflection of the fifth constituent, via the unification variable *$n* which does not agree with any other constituent.

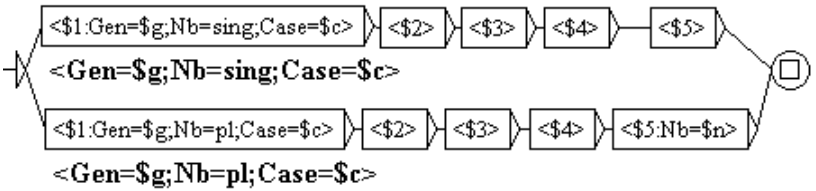

Fig. 19. Inflection graph for *głodny jak wilk* in Polish.

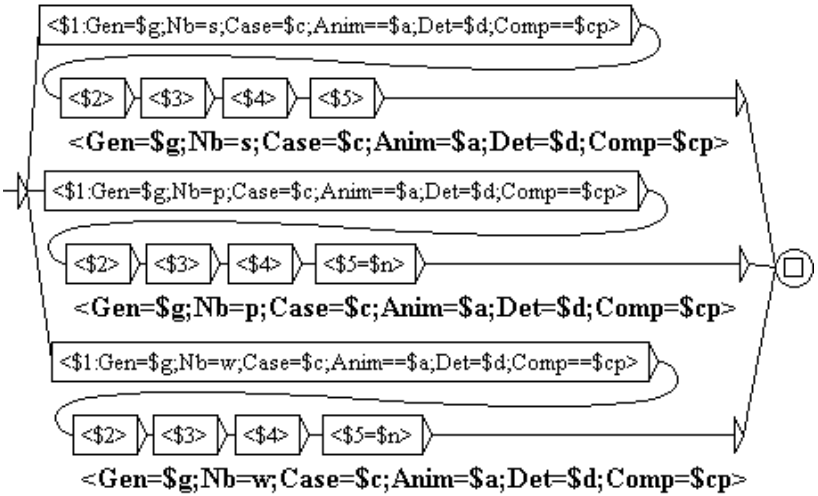

Fig. 20. Inflection graph for *gladan kao vuk* in Serbian.

4.5. Special case of coordination

Some well known problems concern many conjunctive noun phrases which are not necessarily set expressions. Identifying the head word, the inflection features, and the inflected forms of such phrases is not always straightforward. Many coordinative compounds may be seen as containing no headwords, because they are in the plural, although their constituent nouns are in the singular, as in:

(40) (PL) *Adam i Ewa zostali wygnani z raju* ('Adam and Eve were expelled from paradise')

Thus, the compound agrees with the first component noun in case and gender (in Fig. 21, the variables $*g* and *$c* are common to *$1* and to the whole compound) but not in number (*$n1* vs. *pl*).

On the other hand, the compound:

(41) (SR) *alfa i omega, alfom i omegom*, etc. ('alpha and omega')

is feminine singular, thus *alfa* is its headword. The second noun (*$5*) may not agree with the headword in gender (here: *omega* is neuter), number and

animateness, but it does agree in case. Thus, in Fig. 22 the variables *$g5*, *$n5* and *$a5* are specific to *$5*, while *$c* is common to *$1* and *$5*.

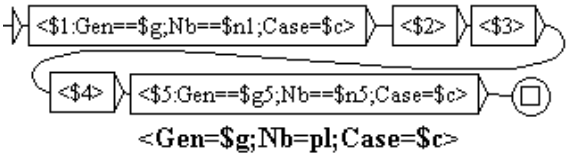

Fig. 21. Inflection graph for *Adam i Ewa* in Polish.

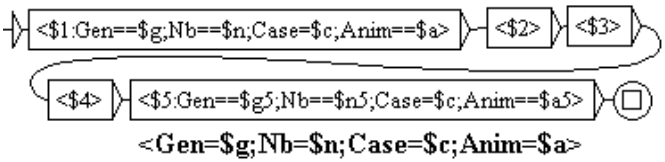

Fig. 22. Inflection graph for *alfa i omega* in Serbian.

Both (40) and (41) are to be seen as irregular, the former due to having no headword, and the latter due to the inflection of the non head constituent.

5. Inflection and variation

According to [Jacquemin 2001] and [Savary, Jacquemin 2003], inflected forms of compounds belong to a more general phenomenon of terminological variation. Variants are two different sequences of lexemes representing the same concept. Variants may be of different nature:

- Orthographic, if the sequences differ in spelling, e.g.:

  (42) (SR) *radio aparat* > *radio-aparat*, *radioaparat* ('radio-set')

  (43) (SR) *država-članica* > *država članica* (’member state’)

- Inflectional (of main interest in this study), if the sequences are different inflected forms of the same compound lemma :

  (44) (FR) *bateau-mouche* > *bateaux-mouches* (‘river boat’)

- Syntactic, if the same component words are used within different, semantically equivalent, syntactic structures. The possible syntactic transformations include:

  - omissions:

  (45) (SR) *profesor engleskog jezika > profesor engleskog* ('teacher of the English language')

  - insertions:

  (46) (FR) *moniteur temps réel > moniteur <u>en</u> temps réel* (‘real-time monitor’)

  - coordinations:

  (47) (PL) *Ameryka Północna, Ameryka Południowa > Ameryka Północna i Południowa* (‘North and South America’)

  (48) (SR) *Banovo brdo, Julino brdo > Banovo i Julino brdo* (‘two Belgrade areas’)

  - inversions, particularly frequent in Slavic languages due to a relatively free word order:

  (49) (PL) *bezwzględna większość > większość bezwzględna* ('absolute majority')

  (50) (SR) *kvadratni metar > metar kvadratni* (‘square metre’)

  - derivational transformations:

  (51) (FR) *tension des artères > tension artérielle* ('blood pressure’)

  (52) (SR) *most na Dunavu > dunavski most* (‘Danube bridge’)

  (53) (PL) *podatek dochodowy > podatek od dochodu* (‘income tax’)

- Semantic, if one or more component words are replaced by their semantic equivalents:

  (54) (FR) *maladie héréditaire > maladie génétique* (‘hereditary disease’)

  (55) (PL) *większość bezwzględna > większość absolutna* ('absolute majority')

- Abbreviation variants, if component words are replaced by their abbreviated forms:

  (56) (FR) *Organisation des Nations Unies* > *ONU* (‘United Nations’)

  (57) (SR) *popularna kultura* > *pop- kultura* (‘pop-culture’)

Different types of variation may combine, thus forming orthographic-and-inflectional variants:

(58) (SR) *država-članica* > *država<u>ma</u> članica<u>ma</u>* (‘member state’ in the singular nominative, and in the plural dative)

inflectional-and-syntactic variants:

(59) (SR) *ministar za unutrašnje poslove* > *ministar unutrašnjih poslova* (‘minister of internal affairs’)

syntactico-semantic variants:

(60) (PL) *większość bezwzględna* > *absolutna większość* ('absolute majority')

abbreviation-and-syntactic variants, etc.:

(61) (SR) *ulje za loženje* > *lož-ulje* ('heating oil')

If all of the above variants are represented within a common framework, their conflation may be very efficiently applied in natural language processing applications, such as information extraction (cf [Jacquemin 2001]).

Within *Multiflex*, orthographic and inflectional variants may be easily expressed. For instance, example (43) can be described by Fig. 23, in which the separator change is expressed by two parallel boxes.

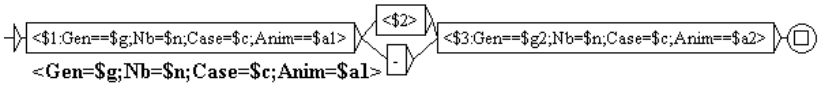

Fig. 23. Representing orthographic and inflectional variants *država-članica* in Serbian.

Omissions, insertions and inversions are also easy to represent by adding new constituents (Fig. 24), by inverting the order of boxes (Fig. 25) within a

path, as well as by bypassing a part of a path and changing the inflection of constituents (Fig. 26).

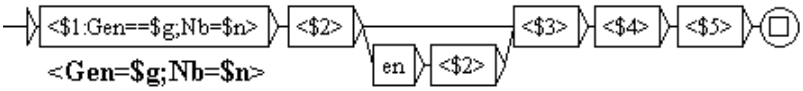

Fig. 24. Insertion variants in *moniteur temps réel* in French.

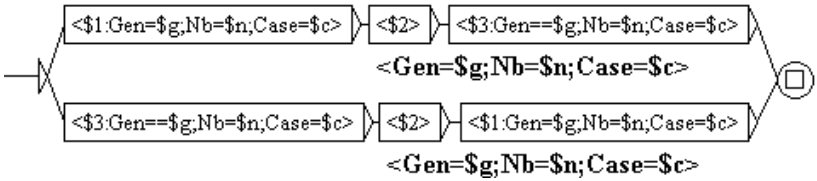

Fig. 25. Inversion variants in *bezwzględna większość* in Polish.

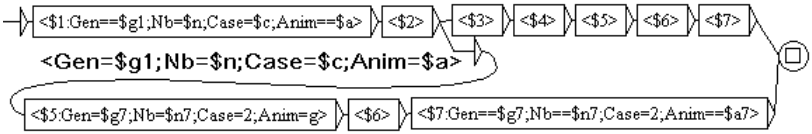

Fig. 26. Syntactic variants in *ministar za unutrašnje poslove* in Serbian.

Other syntactic, semantic and abbreviation variants may be expressed provided that the underlying module for the treatment of simple words allows to express derivations, synonyms and abbreviations within the same framework as inflectional morphology. We haven't experimented with any such module yet.

6. Perspectives

An important ergonomic improvement in the *Multiflex* formalism should affect in future the size of an inflection graph. Presently, a path contains usually as many boxes as there are constituents in a compound. Thus, a graph can easily become unreadable and hard to maintain.

Moreover, a graph can never be shared by compounds of different lengths, such as:

(62) (FR) *réseau à satellites* (‘satellite network’)

(63) (FR) *réseau de transit à satellites* (‘satellite transit network’)

(64) (FR) *réseau de transit haut débit à satellites* (‘high-speed satellite transit network’)

However, all three examples basically inflect in the same way: in order to obtain the plural form one needs to inflect the head noun *réseau* only, regardless of the length of the prepositional phrase that follows. A more compact representation of the sequences of uninflected constituents (e.g. via bounded loops) could allow the three examples above to share the same inflection graph.

Another important perspective with respect to variation would be the possibility of describing non contiguous compound units. In most languages that issue becomes essential if verbal expressions are to be taken into account. However, compound nouns too may be split in a text by external components, as in:

(65) (SR) *žuti karton > Odlazak privrednika u emigraciju žuti je karton poljskim vladama* (‘yellow card’ > ‘…yellow is card …’)

For the time being *Multiflex* does not allow one to account for such kinds of occurrences.

7. Conclusions

We have addressed the inflectional morphology of compounds and other multi-word units in French, Polish and Serbian. Their idiosyncratic behavior requires a descriptive formalism which is at least partly lexicalized. We have presented a previously developed formalism of the *Multiflex* system, which is fully lexicalized.

The major part of our study has dealt with the different linguistic properties of compounds with respect to inflection. We have paid close attention to their morphological non compositionality (exocentric compounds, exceptional formation of some inflected forms, etc.), as well as their orthographic, inflectional, syntactic and semantic variation. We have shown

how most of the properties studied can be described exhaustively and correctly within the *Multiflex* framework.

References

ANSCOMBRE J.-Cl. (1990), Pourquoi un moulin à vent n'est pas un ventilateur, in: *Sur les compléments circonstanciels, Langue Française*, N° 86, Paris, Larousse.

BAUER L. (1983), *English Word-Formation*, Cambridge University Press, Cambridge.

BENVENISTE E. (1974), Fondements syntaxiques de la composition nominale, and Formes nouvelles de la composition nominale, in: *Problèmes de linguistique générale*, vol. 2, Gallimard, Paris, pp. 145-176.

CADIOT P. (1992), À entre deux noms : vers la composition nominale, in: *Lexique*, N° 11, P. U. L., pp. 193-240.

CORBIN D. (1992), Hypothèses sur les frontières de la composition nominale, in: *Cahiers de grammaire*, N° 17, Université de Toulouse Le Mirail, pp. 26-55.

DOWNING P. (1977), On the Creation and Use of English Compound Nouns, in: *Language*, N° 53(4), Linguistic Society of America.

GOUSSIER M. Inflection of compound words, in: *Morphology Recipes, a Cookbook for Building Electronic Lexicons*, Collective volume under a fictive name Marc Goussier, in memory of Maurice Gross, to appear in Oxford University Press.

GROSS G. (1990), Définition des noms composés dans un lexique-grammaire, in: (Courtois, B., Silberztein M. eds.) *Langue Française*, N° 87. *Dictionnaires électroniques du français, septembre 1990*, Larousse, Paris, pp. 84-90.

GROSS G. (1996), *Les expressions figées en français. Noms composés et autres locutions*, Ophrys, Paris.

HABERT B., JACQUEMIN, Ch. (1993), Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques, in: *Traitement Automatique des Langues*, N°2, Hermès-Lavoisier, pp. 5-41. lieu ?

JACQUEMIN Ch. (2001), *Spotting and Discovering Terms through Natural Language Processing*, MIT Press. Lieu ?

JASSEM K. (1996), *Elektroniczny słownik dwujęzyczny w automatycznym tłumaczeniu tekstu*, PhD thesis, Uniwersytet Adama Mickiewicza. Poznań.

KRSTEV C., VITAS D., SAVARY A., (2006), Prerequisites for a Comprehensive Dictionary of Serbian Compounds, *Proceedings of FINTAL'06*, in: (Salakoski T., Ginter F., Pyysalo, S. and Pahikkala T. eds.) LNCS N°4139, Springer, pp. 552-563.

LEVI J. (1978), *The Syntax and Semantics of Complex Nominals*, Academic Press, New York – London.

PAUMIER S. (2002), *Manuel d'utilisation du logiciel Unitex*, http://www-igm.univ-mlv.fr/unitex/manuelunitex.ps accessed on May 25, 2007.

PRZEPIÓRKOWSKI A. (2003). *A Hierarchy of Polish Genders*, in: (Bański M., Przepiórkowski A. eds.) *Generative Linguistics in Poland: Morphosyntactic Investigations*, pp. 109-122.

PRZEPIÓRKOWSKI A. (2004), *The IPI PAN Corpus, preliminary version*, Institute of Computer Science, Warsaw, 91 p., ISBN 83-910948-8-X.

SAVARY A. (2005a), A formalism for the computational morphology of multi-word units, in : *Archives of Control Sciences*, N° 15(3), pp. 437-449. Editeur, lieu?

SAVARY A. (2005b), *MULTIFLEX. User's Manual and Technical Documentation. Version 1.0*, Technical report N° 285, LI-François Rabelais University of Tours, France, 34 p.

SAVARY A., JACQUEMIN Ch., (2003), Reducing Information Variation in Text, in: *LNCS N° 2705*, pp. 145-181, Springer.

SILBERZTEIN M. (1993a), Les groupes nominaux productifs et les noms composés lexicalisés, in: (Chevalier, J.-C., Gross, M., Leclère, Ch. eds.) *Lingvisticae Investigationes, tome XVII:2*, John Benjamins B.V., Amsterdam.

SILBERZTEIN M. (1993b), *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*, Masson, Paris.

SZPAKOWICZ S. (1986), *Formalny opis składniowy zdań polskich*, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

VETULANI Z., WALCZAK B., OBREBSKI T., VETULANI G., (1998), *Jednoznaczne kodowanie fleksji rzeczownika polskiego i jego zastosowanie w słownikach elektronicznych - format POLEX*, Wydawnictwa Naukowe Uniwersytetu Adama Mickiewicza, Poznań.

VITAS D., KRSTEV C. (2001), INTEX and Slavonic Morphology, in: *Actes des 4es Journées INITEX, Bordeaux'01* (Muller C., Royauté J., Silberztein M. eds.), Presses Universitaires de Franche-Comté, pp. 249—263.

WOLIŃSKI M. (2001), Rodzajów w polszczyźnie jest osiem, in : (Gruszczyński W., Andrejewicz U., Bańko M., Kopcińska D., eds.) *Nie bez znaczenia.…*, Wydawnictwo Uniwersytetu Białostockiego, Białystok, pp. 303-305.