Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted July 2008

Computational Inflection of Multi-Word Units

A contrastive study of lexical approaches

Agata Savary

Published by CSLI Publications

Computational Inflection of Multi-Word Units

A contrastive study of lexical approaches

AGATA SAVARY, *Université François-Rabelais de Tours, France*

Abstract

Similarly to simple words, compounds and other multi-word units (MWUs) are subject to inflection. A correct and exhaustive treatment of this issue has an important impact on natural language applications. However it raises some nontrivial questions such as: the role of separators in MWUs, morphological non-compositionality of MWUs, their syntactic and semantic variation, huge sizes of inflection paradigms in highly inflected languages, etc. Due to such problems, the inflectional description of MWUs must be, at least partly, lexicalized. We present a comparative review of eleven lexical approaches to this issue, with respect to linguistic properties of those units. The review is based on case studies of several natural languages. It allows us to put forward some recommendations for a cross-language standard morphological description of MWUs.

LiLT Volume 1, Issue 2, July 2008.
Computational Inflection of Multi-Word Units.
Copyright © 2008, CSLI Publications.

1 Introduction

As shown by Habert and Jacquemin (1993), multi-word units encompass a number of hard-to-define and controversial linguistic objects: compounds, complex terms, multi-word named entities, multi-word lexemes and expressions, collocations, frozen expressions, etc. They may be contiguous or non-contiguous, compositional or non-compositional sequences of words, and may admit graphical, morphological, syntactic and semantic variation. Numerous linguistic and pragmatic definitions of compounds and other MWUs (Benveniste (1974), Downing (1977), Levi (1978), Bauer (1983), Gross (1990), Anscombre (1990), Corbin (1992), Cadiot (1992), Silberztein (1993b), Gross (1996), Sag et al. (2002), etc.) invoke three major points:

- they are composed of two or more graphical words
- they show some degree of morphological, syntactic, distributional or semantic non-compositionality
- they have unique and constant references

However, the basic notions (a word, a reference, the non-compositionality) and measures (degree of non-compositionality), used in those definitions are themselves controversial. For instance, as shown below, the notion of a graphical word may be application-dependent and/or language-dependent. Thus (in accordance with the approaches whose comparative study we present below), we consider a MWU as a sequence of graphical units which, for some application-dependent reasons, has to be listed, described and processed as a unit. In most cases the graphical units composing a MWU are themselves morphologically analyzable. A broader discussion on how to define a MWU is out of this paper's scope.

The quantitative and qualitative importance of multi-word units in natural languages is now widely acknowledged. They are placed on the frontier between morphology and syntax because of their hybrid nature: some of their properties are idiosyncratic (which suggests a lexicalized description), while some others are productive (which is more easily reflected by a grammar). In this study we are particularly interested in the inflectional properties of MWUs, which are however often connected to phenomena on the graphical, syntactic and semantic level.

Obviously, a reliable inflection processing of single words is a necessary condition for the inflection processing of MWUs. However, this condition is rarely a sufficient one. For example, in order to obtain the plural form of *chief justice* and *lord justice* in English not only do we need to know how to generate the plural of *chief, lord* and *justice* but also to know how different inflected forms of these constituents combine. For instance the following plural forms are correct:

(1) *chief justices*

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

but not **chiefs justice* and **chiefs justices*. There are however few automatically accessible hints indicating that the former compound is morphologically a standard English *Noun-Noun* phrase taking an *s* at its last constituent in plural, while the plural of the latter one has three variants. Obviously, some lexicalized description is needed in order to account for this idiosyncratic behavior.

A correct and exhaustive inflectional analysis and generation of MWUs is one of the conditions for a high-quality natural language application. Studies concerning automatic treatment of MWUs have been performed for two decades. Some in-depth linguistic and computational approaches to word composition, aiming at general language modeling, have co-existed with numerous robust statistical methods, sometimes augmented with some linguistic knowledge. Nowadays, there is a growing conviction in the NLP community that large linguistic lexicons and grammars of MWUs are needed, due to their two characteristics: (i) they represent a high percentage of items in natural language corpora, (ii) most of them, taken separately, appear very rarely in corpora. For instance, Gross and Senellart (1998) showed that more than 40% of all tokens in a one-year corpus of the French journal *Le monde* belong to multiword units or expressions, and should not be analysed individually. Savary (2000) proved that 85% of all graphically distinct compound noun forms appear less than twenty times in a one-year corpus of the *Herald Tribune*. Baldwin and Villavicencio (2002) experimented with a random sample of two hundred English verb-particle constructions and showed that as many as two thirds of them appear at most three times in the Wall Street Journal corpus. Sag et al. (2002) cite some studies considering the number of multi-word expressions as high as the one of single words, and argue that these figures are an underestimate, especially in terminological sublanguages.

The main aim of our study is analyzing the state of the art in the lexicon-oriented computational treatment of the (largely understood) inflectional morphology of MWUs. This paper is organized as follows. In section 2 we perform a study of linguistic properties of MWUs, with a particular focus on inflection, which we illustrate with examples in English (EN), French (FR), Polish (PL), Serbian (SR), German (DE) and Turkish (TU). In section 3 we study eleven existing lexical approaches to MWUs inflection in several natural languages. In section 4 we compare these approaches with respect to how well they account for the linguistic properties shown. In section 5 we conclude with some recommendations concerning cross-language universal lexicalized description of MWUs.

2 Linguistic study of MWUs' inflection

2.1 Graphical aspects

As mentioned above, the formal definition of a multi-word unit is a controversial issue. One of the points that remain unclear is how to distinguish MWUs from simple words on the one hand, and from the so-called "free sequences" on the other hand, on the purely graphical level. This distinction is particularly important in computational applications, where text tokenization (cf. Grefenstette and Tapanainen, 1994), i.e. its division into graphical tokens, is most ofter the prerequisite for further automatic processing. The role of separators is crucial for this problem. In some approaches, such as Silberztein (1993a), a lexical unit is a contiguous sequence of alphabet characters, while a MWU is, from the graphical point of view, a sequence of at least two graphical units, separated by non-alphabet characters. Thus, the sequence (FR) *aujourd'hui* ('today') is seen as a MWU consisting of two graphical units, *aujourd* and *hui*, separated by an apostrophe, while in the English contraction *don't* the last character is seen as a token on its own. In other approaches, some punctuation marks, such as apostrophe and hyphen, are allowed to be inherent members of simple words. Thus, *aujourd'hui* and *don't* are considered as simple words.

Separators as constituents

The role of non-alphabet characters is however not limited to separating two components. If an alphabet is defined as a language-dependent closed list of letters, e.g. the twenty-six Latin letters in English, then characters not belonging to this list may still be inherent members of compounds with a genuine semantic content, as in the following examples:

(3) (EN) *λ-calculus*

(4) (FR) *rayon γ* ('gamma ray')

(5) (PL) *Windowsy 3.11, Windowsów 3.11, etc.* ('name and version of an operating system', in nominative, genitive, etc.)

Moreover, separators may be either disambiguating elements or source of orthographic variants. For instance, the occurrence of (6) in a text is clearly a compound, while (7) is probably a sequence on the frontier of two different phrases, as in *put the hanger on the floor*. Conversely, the presence and the absence of separators is allowed, often in unpredictable manner, within some lexicalized compounds, such as (8):

(6) (EN) *hanger-on*

(7) (EN) **hanger on*

(8) (SR) *radio aparat, radio-aparat, radioaparat* ('radio-set')

Finally, separators may be relevant to inflection, as in:

(9) (PL) *PZPR* (nomin.), *PZPR-u* (gen.), *PZPR-owi* (dat.), etc.

(10) (PL) *Sartre* (nomin.), *Sartre'a* (gen.), *Sartre'owi* (dat.), etc.

Squeezed MWUs

On the other hand, the obligatory presence of separators within MWUs is questionable because some contiguous sequences of letters behave morphologically as compounds:

(11) (EN) *passerby, passersby*

(12) (DE) *Schul|kind, Schul|jahr, Schul|lehrer,* ... ('pupil, school year, teacher')

(13) (FR) *bon|homme, bons|hommes* ('fellow')

(14) (PL) *chciał|bym, chciała|bym* ('I would like, in masculine and feminine')

2.2 Morphosyntactic compositionality

The *compositionality* of a compound means that its various (morphological, distributional, syntactic and/or semantic, etc.) properties can be fully deduced from the respective properties of its constituents. In the scope of our study the *inflectional* compositionality of compounds is of the main interest. It is closely connected to the linguistic notion of the *head word*, i.e. the constituent whose morphological properties determine those of the whole compound. For instance, the phrase:

(15) (PL) *Polska Akademia Nauk, Polskiej Akademii Nauk*, etc. ('Polish Academy of Sciences' in nominative, genitive, etc.)

is a noun in feminine singular nominative and genitive, as its underlined head word *Akademia*/*Akademii* is. Moreover, in such *Adjective Noun* $Noun_{genitive}$ structures, typical for Polish compounds, the non-head components may be affected by agreement and government rules imposed by the head word. Here, *Polska* has to agree in gender, number and case with *Akademia*, while *Nauk* remains always in feminine genitive plural.

If all compound phrases to be recognized were perfectly compositional in this sense, the description of their inflection could be done by: (1) a simple words' lexicon describing the properties of constituents, and (2) a general phrase grammar allowing to derive the properties of compounds from those of their constituents. Many such *phrase grammars* have been created for different languages. Their major advantage is to factorize the description of general inflectional phenomena and thus to avoid systematic entry-per-entry treatment. Their important drawback is that they may incorrectly treat exceptional inflection (cf. section 2.3) or compound structures that are morphosyntactically ambiguous as in:

(16) (EN) *man servant, men servants*

(17) (EN) *man eater, man eaters*

Thus, in a fully correct and exhaustive approach, a MWU's inflection should be accounted for, at least partly, at the lexical rather than grammatical level.

2.3 Morphosyntactic non-compositionality

It is well known that in most cases compounds are, at least partly, non-compositional. This fact is considered as a defining criterion of compounds with respect to free structures in Gross (1988). The *inflectional* non-compositionality may be observed in several cases discussed below. A more detailed study of this phenomenon in French, Polish and Serbian has been performed by Savary et al. (2007).

Exocentric MWUs

A phrase is exocentric if it has no headword, i.e. it contains no word from which its inflectional properties might be deduced, as in the following structures:

(18) (FR) *un perce-neige, des perce-neige* (literally: 'pierce-snow' = 'snowdrop')

(19) (FR) *un perce-oreille, des perce-oreilles* (literally: 'pierce-ear' = 'earwig')

(20) (EN) *a drive-in, drive-ins*

(21) (EN) *a four-in-hand, fours-in-hand, four-in-hands* ('coach pulled by four horses and driven by one person')

(22) (EN) *attorney general, attorney generals, attorneys general*

In the two former examples, *perce* is a genderless verb form, *neige* and *oreille* are feminine, while the compounds themselves are masculine. In the two latter ones *in* and *four* cannot be considered as regular headwords because, as individual words, they don't admit plural. In the last example, if any of the two nouns were the headword, it would always have to agree in number with the whole compound, which is not the case.

Agreement Irregularities

As said before, in perfectly compositional MWUs the morphosyntactic structure of the multi-word lemma determines the agreement and government rules imposed by the headword. These rules may be defied in three kind of situations:

- An agreement does not occur when it normally should. For instance, the compound noun:

  (23) (FR) *grand-mère, grand-mères, grands-mères* ('grandmother')

is of a typical *Adjective-Noun* structure, in which the two constituents agree in number and gender. However *grand* is always masculine while *mère* is feminine. Moreover, *grand* may or may not respect the number agreement.

- An agreement occurs when it normally shouldn't.

  (24) (FR) *toile d'araignée - toiles d'araignée, toiles d'araignées* ('a spider's web')

  This compound is of a standard *Noun de Noun* construction, in which typically only the first noun inflects. Here however, two plural variants are admitted in which the second noun may or may not carry the inflection mark. Analogous examples in English are (2) and (16).

- An agreement should or shouldn't occur, and it occurs partially.

  (25) (FR) *bateau mouche, bateaux mouches* (literally 'a fly boat' = 'a Paris-style river boat')

  (26) (PL) *majster klepka, majstra klepki, etc.* (literally 'master floorboard' = 'an incompetent')

  Both compounds are appositions, i.e. *Noun Noun* constructions, in which the necessity of agreement between the two nouns is unclear. Supposing that both nouns should typically agree, the two above examples are irregular because the head noun *bateau* and *majster* are masculine while the two other nouns, *mouche* and *klepka* remain always feminine. If, conversely, the agreement within appositions is not required, these compounds are also irregular because their constituents do agree in number, and number and case, respectively. Yet another hypothesis saying that nouns in appositions typically agree in number and case (if relevant) but not in gender, is defied by examples like (27) in which both nouns fully agree.

  (27) (FR) *assistant approvisionneur, assistants approvisionneurs, assistante approvisionneuse, assistantes approvisionneuses*
  ('assistant provisioner' in masc. sing., masc. pl., fem. sing. and fem. pl.)

  Such irregularities can only be solved either by lexicalized description, or by redefining the inflection categories according to the inflectional behavior of words, as in Przepiórkowski and Woliński (2003). Thus, the traditional category of nouns should be divided into two subcategories: (i) nouns having a fixed gender, such as *mouche*, (ii) nouns inflecting in gender, such as *approvisionneur*. With this distinction, general, i.e. non-lexicalized, grammar rules could capture the agreement particularity of (25) with respect to (27).

Defective Inflection Paradigms

In some MWUs at least one inflected form that is usually expected for the structure concerned is inexistent. For instance, the compounds:

(28) (EN) *bits and pieces, *a bit and a piece*

(29) (PL) *zimne nogi, zimnych nóg, . . . , *zimna noga* (literally 'cold legs'='a dish consisting of meat and jelly', in nominative plural, genitive plural, etc.)

do not admit singular forms, even if *a bit and a piece*, as well as *zimna noga, zimnej nogi*, etc., are syntactically correct sequences (in singular these phrases loose their particular sense). Note that the above examples differ from the ones whose inflection is fixed but not defective, such as *cross-roads*:

(30) (EN) *The bits and pieces he usually kept in his pocket were now on the table.*

(31) (EN) **The bits and pieces he usually kept in his pocket is now on the table.*

(32) (EN) *All cross-roads in the main street were blocked by the police.*

(33) (EN) *The cross-roads in front of my house was blocked due to an accident.*

Note also that the non-existence 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 following compounds:

(34) (EN) *security police*

(35) (FR) *funerailles nationales* ('national funeral')

(36) (PL) *krótkie spodnie* ('shorts')

do not admit a singular form due to the fact that their head nouns *police*, *funerailles* and *spodnie* are themselves plural-only nouns.

2.4 Inflection and variation

According to Savary and Jacquemin (2003), inflected forms of compounds belong to a more general phenomenon of terminological variation. In particular, variants may result from separator alternation, as in (8), as well as a large range of other linguistic transformations:

- Insertions:

  (37) (FR) *moniteur temps réel, moniteur en temps réel* ('real-time monitor')

- Omissions:

  (38) (SR) *profesor engleskog jezika, profesor engleskog* ('teacher of the English language')

- Order change:

(39) (PL) *bezwzględna większość, większość bezwzględna* ('absolute majority')

- Duplications:

(40) (TU) *ev* ('house', in Turkish), *ev ev* ('house by house')

- Derivational transformations:

(41) (FR) *tension des artères, tension artérielle* ('blood pressure')

- Semantically motivated replacements:

(42) (FR) *maladie héréditaire, maladie génétique* ('hereditary/genetic disease')

- Abbreviations:

(43) (EN) *physical education, phys-ed*

(44) (EN) *United Nations, UN*

Orthographic, inflectional, syntactic and semantic variants may exist side by side, as in:

(45) (EN) *student union, students union, students' union*

(46) (EN) *birth date, date of birth*

(47) (SR) *ministar za unutrašnje poslove, ministar unutrašnjih poslova* ('minister of internal affairs')

2.5 Inflectional paradigm and base form

The inflectional paradigm of a MWU is the list of its inflected forms together with their inflectional description. The size and contents of this list depend clearly on the nature of the language studied (e.g. French adjectives usually have four inflected forms while Polish ones have several dozens of them). However, they also depend on the morphological model chosen for the given language. Firstly, it is not always obvious how to tell the inflectional from the derivational morphology. For instance, the past participle form of a French verb (e.g. *voir* → *vu*), is usually seen as its inflected form, but this form admits itself an adjectival inflection. Thus, the question is if the inflected forms of the past participle (*vu*, *vus*, *vue* and *vues*) should or should not belong to the inflection paradigm of the verb. Secondly, it is sometimes unclear how to determine the precise list of the possible inflection values (singular, plural, feminine, etc.). For instance different approaches estimate the number of Polish genders at five, six, eight, or eleven, respectively. Thus, the corresponding inflection paradigms of adjectives (that inflect in gender, number and case) may contain up to 70, 84, 110, or 154 forms (many of them syncretic).
The large size of an inflection paradigm in highly inflected languages, such as Slavic or concatenative languages, is a problem as such in a formal lexical

approach. For the sake of human efficiency large numbers of forms should be describable by compact rules. At the same time the formalism should be precise enough to avoid overgeneralization and overlooking of exceptions. See appendix 2.4 to appreciate the size of an inflectional paradigm of a Serbian compound noun.

The notion of a base form is essential in the morphological analysis and generation of inflected forms. It may be seen either as the canonical representative of the inflection paradigm, or just its identifier. In the first case the base form belongs itself to the paradigm (i.e. it is a linguistically correct form, called a *lemma*). In the second case it may well be an abstract (linguistically incorrect) form. Consider for instance:

(48) (EN) *customs barrier, customs barriers*

(49) (FR) *mémoire vive, mémoires vives* (literally 'live memory'='random access memory')

where *mémoire* is a feminine noun and *vive* is the feminine form of the adjective *vif*. These sets of compound forms may be represented either by their first elements or by "abstract" forms *custom barrier* and *mémoire vif*. For an efficient usage and treatment of MWUs by humans (e.g. consulting MWU lexicons, or validation of automatically extracted candidate terms), the former solution is more appropriate.

2.6 Noncontiguous MWUs

Multi-word expressions (MWEs), particularly those containing verbs, are MWUs which may appear in the corpus as noncontiguous sequences of items, as in:

(50) (EN) *He has finally made up his bloody mind.* (the MWE's components are underlined)

An exhaustive description of such expressions remains a challenge. Their precise analysis is out of the scope of this paper but we will pay attention to how the approaches presented below provide, at least partly, a framework for this phenomenon.

3 Lexical approaches to the inflection of MWUs

Due to the morphosyntactic non-compositionality of many MWUs, as well as their semantic opacity, studies on their lexicalized description have been performed for two decades. The variety of linguistic and computational approaches in this domain is comparable to the number of those proposed for the morphology of simple words. In this section we present a review of some of these lexicalized approaches to the inflection MWUs, in view of their comparative analysis, and best-practice recommendations.

3.1 DELA dictionaries

In the Paris school of DELA electronic dictionaries (see Courtois and Silberztein, 1990) simple and compound words are systematically listed and their inflectional description is done on the entry-per-entry basis. In a so-called DELAS[1] lexicon the inflectional paradigm of each simple word is described by an *inflectional code* representing a set of sequences of operators applied to the word's lemma (such as 'cut the last symbol', 'add a new symbol', 'move one symbol to the left', etc.), together with the corresponding inflectional features attached to each form produced (such as 'feminine plural', etc.). For instance, example (51) represents one DELAS entry, *cousin* attached to the inflection code *N32*, depicted in (52). The code states that the masculine singular form of a word is equivalent to its lemma (the '-' sign means no operator), the feminine singular is produced by adding suffix *e* to the lemma, etc., which results in the set of four inflected forms of *cousin* shown in (53). The application of inflection codes to all entries in the DELAS allows for an automatic construction of a DELAF[2], which can be compressed into a finite-state tool before being applied in the process of morphological analysis.

(51) *cousin* ⇒ code: *N32*

(52) *N32: (-/⟨masc,sing⟩,s/⟨masc,pl⟩,e/⟨fem,sing⟩,es/⟨fem,pl⟩)*[3]

(53) *cousin* ⇒ lemma: *cousin*, gender: *masc*, number: *sing*

cousins ⇒ lemma: *cousin*, gender: *masc*, number: *pl*

cousine ⇒ lemma: *cousin*, gender: *fem*, number: *sing*

cousines ⇒ lemma: *cousin*, gender: *fem*, number: *pl*

The inflectional description of compounds has found several solutions in this school, some of which we present below. In each case, only contiguous compounds have been described. Their lemmas have been listed and described in

[1] DELAS stands for the *LADL's electronic dictionary for simple words*, where the LADL is the central laboratory having proposed the methodology.

[2] LADL's electronic dictionary for inflected forms of simple words

[3] The precise formalism is somewhat different but equivalent to what is shown in this example.

a controlled way within DELAC[4] dictionaries, which, similarly to the DELAS, can be automatically inflected into DELACF[5] dictionaries, and compressed before being applied to texts. Syntactic properties of simple and compound units and expressions are further described by *lexicon grammars*. In sections 3.1, 3.1, 3.1 and 3.1 we propose an abstract graphical representation of DELAC and DELACF entries. See the appendix for detailed formats admitted in each approach.

The DELA methodology addresses the morphological analysis of compounds in an extensive manner, i.e. by matching word sequences in text against a full (compressed) list of compound inflected forms (a DELACF) obtained automatically from compound annotated lemmas (a DELAC).

French DELAC

The French DELAC was the first to be created. It contains 126,000 compound lemmas which yield 271,000 inflected forms. In Silberztein (1993a), the inflectional paradigms for compounds, unlike those for simple words, are not designated by autonomous inflectional codes. They are created instead by *ad hoc* language-dependent filters with a manual post-filtering of ambiguities and exceptions. For instance, in the DELAC entry (54), *N32* and *A32* are the inflectional codes of *cousin* and *germain*, respectively, the whole compound is masculine singular, and it admits gender and number inflection.

(54)

cousin	*germain*	⇒
code: *N32*	code: *A32*	gender: *masc* number: *sing* gender inflection: *yes* number inflection: *yes*

The automatic inflection of a lemma consists in inflecting all constituents to which a code has been attributed (here: *cousin* and *germain*), in all inflection categories admitted (here: gender and number), and imposing that these constituents agree. Here, the corresponding DELACF entries obtained are 55. Each form is attached to its lemma and its morphological description.

(55)

cousin germain	⇒	lemma: *cousin germain* gender: *masc* number: *sing*
cousins germains	⇒	lemma: *cousin germain* gender: *masc* number: *pl*

[4]LADL's electronic dictionary for compounds
[5]LADL's electronic dictionary for inflected forms of compounds

cousine germaine ⇒ lemma: *cousin germain*; gender: *fem*; number: *sing*

cousines germaines ⇒ lemma: *cousin germain*; gender: *fem*; number: *pl*

Inflection variants, such as in (22) through (24) and (37) through (47), require separate lexicon entries, for instance:

(56) *toile* d'araignée; code: *N21* ⇒ gender: *fem*; number: *sing*; gender inflection: *no*; number inflection: *yes*

toiles d'araignées ⇒ gender: *fem*; number: *pl*; gender inflection: *no*; number inflection: *no*

The first entry yields two DELACF forms:

(57) *toile d'araignée* ⇒ lemma: *toile d'araignée*; gender: *fem*; number: *sing*

toiles d'araignée ⇒ lemma: *toile d'araignée*; gender: *fem*; number: *pl*

while the second one produces the plural variant attached to a different lemma:

(58) *toiles d'araignées* ⇒ lemma: *toiles d'araignées*; gender: *fem*; number: *pl*

Since the inflection codes for simple words may only apply to their lemmas and do not allow to transform any inflected form directly into another inflected form, it is unclear how abstract base forms would be treated in this approach. For instance in example (49) the lemma of the second constituent is the masculine form *vif*. If the compound lemma is:

(59) *mémoire*; code: *N21* — *vif*; code: *A38* ⇒ gender: *fem*; number: *sing*; gender inflection: *no*; number inflection: *yes*

then it is unclear how the rule of making both constituents agree is applied if no gender inflection is allowed. If however the lemma is:

(60)

mémoire	*vive*
code: *N21*	code: *A21*

⇒

gender: *fem*
number: *sing*
gender inflection: *no*
number inflection: *yes*

then the adjective *vive* implies an artificial adjectival lemma having only the feminine forms *vive* and *vives*. Similar doubts apply to exocentric compounds in which default agreement of the inflected constituents is impossible.
The inflection tool accompanying this formalism needs adaptation to the morphological model of each new language if only new inflection categories (e.g. case) or values (e.g. neuter gender) are needed.

English DELAC

In Savary (2000) the English DELAC of 60,000 compounds lemmas and the corresponding DELACF of 110,000 inflected forms are constructed. The previous model is enlarged in that: (i) simple constituents in a compound lemma are annotated by their DELAF-entries, (ii) characteristic constituents, i.e. the headword and the words agreeing with it, are pointed out, (iii) exceptional forms are explicitly described.
Examples (16) and (17) are represented by the following samples:

(61)

man 1	*servant* 2
lemma: *man* code: *N8* number: *sing*	lemma: *servant* code: *N1* number: *sing*

⇒

class: *noun*
number: *sing*
inflection: *{number}*
char. const.: {1, 2}

(62)

man 1	*eater* 2
	lemma: *eater* code: *N1* number: sing

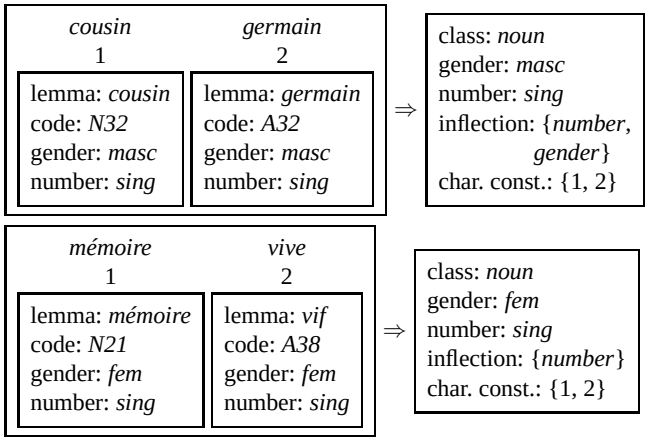
⇒

class: *noun*
number: *sing*
inflection: *{number}*
char. const.: {2}

Each individual constituent obtains an ordinal number indicating its position in the compound, as well as its DELAF label, as in (53), and its inflection code. Here both compounds are nouns in singular admitting the number inflection. However, in (61) both constituents are characteristic (they agree in both numbers), while in (62) only the second one (*eater*) is. The inflection of the compounds is done by default, i.e. by inflecting all characteristic constituents and making them agree (which is trivial in English as only the plural

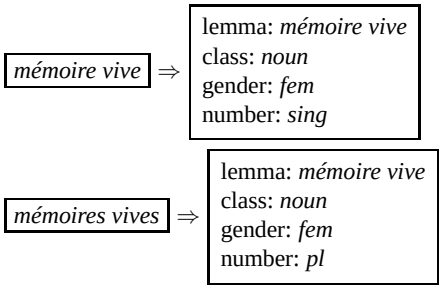
form is concerned). This process yields *man servant* and *man eater* in singular, as well as *men servants* and *man eaters* in plural.
The formalism adapts to a different language via a morphology configuration file specifying the possible inflection classes (number, gender, case, etc.) and their possible values (singular, feminine, nominative, etc.).
Thus, the French DELAC-entries (54) and (60) can be described as follows:

(63)



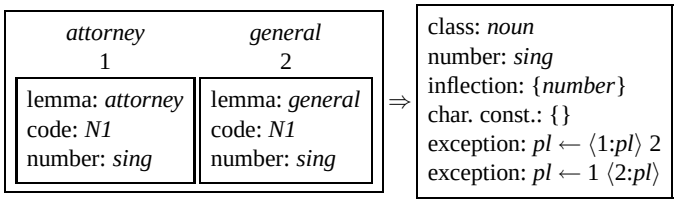
Note that annotating the inflected components with their DELAF-entries allows to avoid both abstract base forms for compounds and artificial lemmas for simple words (cf. examples (59) and (60)). Here, the plural form *vives* is not obtained directly from *vive* but from the attached lemma *vif*. The DELACF entries obtained from this description contain the same non-abstract lemma *mémore vive*:

(64)



Non-compositional and irregular examples, as (18) through (26), are describable, as in the following sample:

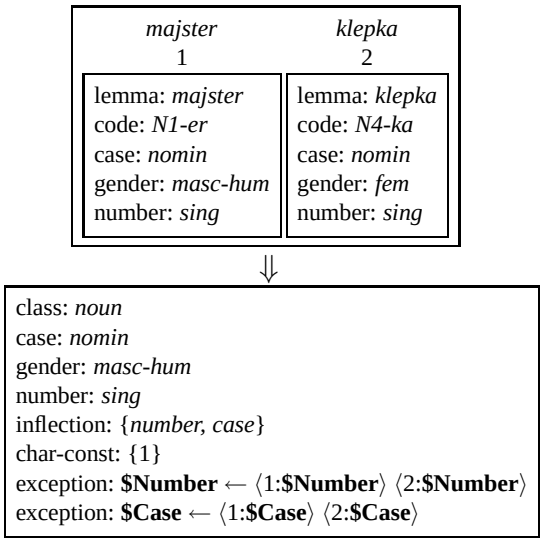
(65)



Here, the entry inflects in number, and has no characteristic constituent (it is exocentric). Thus its plural formation is not done by default (i.e. not by inflecting the characteristic constituents), but follows two exception rules: one needs to inflect the first constituent into plural and leave the second constituent unchanged (*attorneys general*), or conversely (*attorney generals*).

A unification formalism allows to compactly express large paradigms concerned by agreement rules. Thus, example (26) may be represented as follows:

(66)



The exception rules use unification variables *$Number* and *$Case* to indicate that any number and any case of a compound is obtained by inflecting and unifying both constituents, despite the fact that only the first one is characteristic.

As seen in the appendix, section 1.2, one drawback of this formalism is the distribution of the morphological description between the compound entry and the preamble (i.e. lines beginning with % in appendix 1.2 and describing the characteristic constituents and the exception rules). An entry may not be regarded independently from the sublist it appears in. In particular sorting the textual lexicon is not allowed.

The formalism also suffers from the lack of expressive power allowing the attachment of orthographic, syntactic and semantic variants to a common lemma. Thus, examples (37) through (47) need separate entries for each variant.

Greek DELAC

In Kyriacopoulou et al. (2002), the set of all inflected forms of a Greek compound may be obtained by the application of *restriction filters* to the set of all possible combinations of the inflected forms of the particular constituents. For instance in (67), *N33403*, *DET* and *N125* are inflection codes for the three constituents. The third component always remains in genitive singular (see the restriction filter), while the two others may inflect freely.

(67)

κλειδί	*τον*	*παραδεί"σον*
1	2	3
code: *N33403*	code: *DET*	code: *N125*

⇒

'a key to paradise'
class: *noun*
restriction: ⟨3:*genit*, *sing*⟩

In this model, there is a lack of flexibility as to the possible combinations of inflection features admitted by the filters. For instance, cases such as e.g. (22) would be hard to express and would probably require two separate entries. Thirdly, the rules for determination of inflection features of a compound are heterogeneous and non-generic, which makes the formalism hardly adaptable to a different language. For instance, exocentric compounds such as (18) require a graph-based description, in which all inflected forms have to be cited explicitly. Finally, the presence or absence of constituents, or their order changes, cannot be expressed by restriction filters.

NooJ DELAC

In Silberztein (2005) a uniform formalism for the inflection of both compounds and simple words is suggested. The previous DELA operators, applying to simple word lemmas, are completed by compound-oriented operators like 'go to the end of the next word' or 'go to the end of the first constituent'. Thus, any inflected form of a compound results from a sequence of actions performed on the suffixes of particular graphical constituents of an entry. For instance, the following inflection code:

(68) ACTOFGOD = <E>/singular + <PW>s/plural

represents a set of two forms: the singular is created by recopying the lemma without any change ("<E>" represents an empty sequence of operators), while the plural is obtained by going to the end of the first constituent ("<PW>" operator) and adding an *s*. Such a code may apply to the English compounds of different lengths, where the first constituent inflects, such as *act of God*, *balance of payment deficit*, *member of the opposite sex*, etc.

In this formalism non-abstract base forms as well as non-compositional and irregular compounds are describable, because the compositionality is not an issue. A compound is seen as a sequence of characters, including blanks, and no underlying morphological description of simple words is needed. Inflection variants can be described by simply aligning as many operator sequences as needed for a particular form. For instance, compound (22) can be described by the following code (adding an empty suffix produces the singular, going to the end of the first word and adding an *s* produces the plural, adding suffix *s* at the end of the whole compound produces another plural variant):

(69) ATTORNEYGENERAL = <E>/sing + <PW>s/pl + s/pl

This non-compositional approach to inflection of compounds is however also its main problem. The morphology of a simple word has to be described as many times as this word appears in a compound at a position subject to inflection. For example, the description of a compound containing *battle* cannot rely on a unique (thus, easily maintainable) description of *battle* as a simple word but must be repeated for all numerous compounds of types *battle royal*, *battle of nerves*, *running battle*, etc. This problem becomes important in case of highly inflected, e.g. Slavic, languages.

Multiflex

In Savary (2005) a formalism for inflection of compounds, implemented in the *Multiflex* system, allows to gather all the inflectional description of a compound within a graph. Each path in the graph describes one or more inflected forms. A unification mechanism allows to account for agreements within constituents, and to represent huge inflection paradigms compactly. For instance

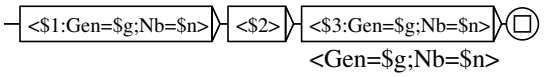
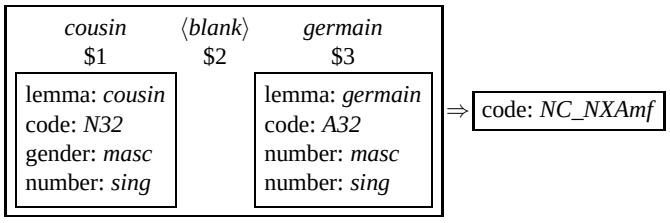


FIGURE 1 *Multiflex* inflection graph NC_NXAmf for *cousin germain*

figure 1 describes the regular French compounds inflecting like *cousin germain* (cf. example (54)). Morphological categories *Gen* and *Nb*, as well as their corresponding morphological values ({*sing, pl*}, {*masc, fem*}, etc.), are language-dependent. The first constituent (*$1*), here *cousin*, inflects in gender (*Gen*) and number (*Nb*). The unification variables assigned to each of these categories (*$g* and *$n*) may take any value of the respective category domains ({*masc,fem*} and {*sing,pl*}, respectively). The second constituent (*$2*), here the blank space, remains unchanged (no operator present in this box). The third constituent, here *germain*, inflects similarly to the first one. The unification variables are common in the first and the third box, which means that

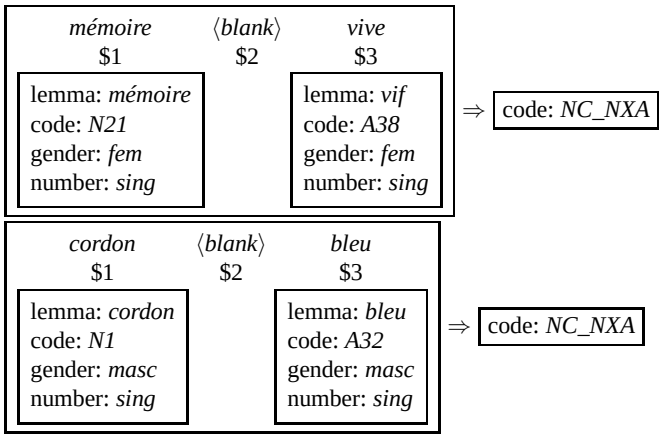
those constituents agree both in gender and in number. The category-value equations present below the third box describe the morphological values of the whole compound. Here the gender and the number of the compound are the same as those of the first and the third constituent.
With this graph, the DELAC entry in (54) and (63) takes the following form:

(70)



and the total exploration of the graph *NC_NXAmf* results in a list of DELACF entries similar to (55).
Note that if the unification mechanism were not available, each of the four inflected forms of lemmas like (70) would have to be described by a separate path in the graph in fig. 1 (the first path imposing the singular masculine form for the first and the third constituent, the second one imposing the masculine plural, etc.). In Slavic languages such method would rapidly turn into a nightmare, with paradigms containing several dozens of forms, each of which would need a separate path in the corresponding graph.
A value inheritance operator allows to assign the same inflection graph to lemmas inflecting similarly but having different inflection values. For instance the graph on figure 2 applies to both entries below:

(71)



despite their different gender. Note that figure 2 differs from 1 only by the double assignment ('==') of variable *$g* to *Gen* in the first box. This operator means that variable *$g* may take only one gender value - the value that the

corresponding constituent has in the compound's lemma. For instance, in the lemma *mémoire vive* shown in (71), the first constituent *mémoire* is in feminine gender. Variable *$g* inherits this gender value and remains unchanged during the whole inflection process. Similarly, in lemma *cordon bleu*, the first constituent *cordon* is in masculine, so when the graph is applied to this compound, variable *$g* inherits this gender value and, again, remains unchanged. The correctly annotated DELACF entries for these examples can be seen in the appendix, section 2.2).

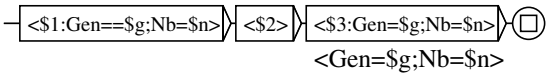
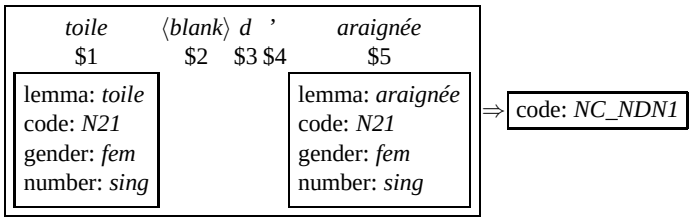


FIGURE 2 *Multiflex* inflection graph NC_NXA for *mémoire vive* and *cordon bleu*

Any combination of morphological values can be expressed in this formalism, thus all non-compositional and irregular examples, such as (18) through (26) can be described. For instance, figure 3 shows the inflection graph for compound (72).

(72)



The gender of the compound is inherited from its first constituent (*$1:Gen==$g*). The fifth constituent may be either left intact (upper path) or it may agree in number with the first constituent (*$5:Nb=$n*). The lack of the category-value equation for the gender of the fifth constituent means that its gender never changes, and does not influence the gender of any other components.

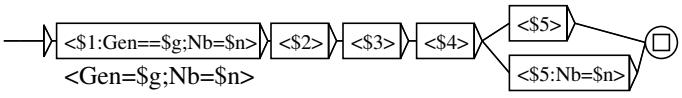


FIGURE 3 *Multiflex* inflection graph NC_NDN1 for *toile d'araignée*

Since individual components of the lemma are referred to via ordinal variables *$1*, *$2*, etc., deletions, insertions, duplications, and order changes of components may be expressed, as in variants (37) through (40), and (45) through (47). For instance, entry (73) and figure 4 describe example (45). The first constituent may be either unchanged, or inflected into plural and

followed by an optional apostrophe. In the resulting DELACF entries (74) all variants are attached to the same lemma.

(73)

student $1	⟨*blank*⟩ $2	*union* $3	⇒
lemma: *student* code: *N1* number: *sing*		lemma: *union* code: *N1* number: *sing*	code: *NC_NXN1*

(74)

student union	⇒	lemma: *student union* class: *noun* number: *sing*
students union	⇒	lemma: *student union* class: *noun* number: *sing*
students' union	⇒	lemma: *student union* class: *noun* number: *sing*
student unions	⇒	lemma: *student union* class: *noun* number: *pl*
students unions	⇒	lemma: *student union* class: *noun* number: *pl*
students' unions	⇒	lemma: *student union* class: *noun* number: *pl*

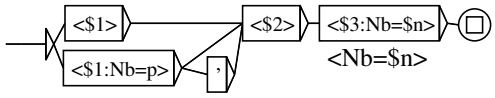


FIGURE 4 *Multiflex* inflection graph *NC_NXN1* for *student union*

The system is designed so as to remain relatively independent from the underlying morphological description of simple words, as long as its model can be described by a list of inflectional categories (gender, number, case, etc.) together with their possible values (singular, plural, masculine, etc.). In particular, non-alphabetic characters can be considered as constituents, and word

boundaries can be freely defined, and are not limited to blanks and punctuation marks, as in examples (3)-(14).
In Krstev et al. (2006a) the system has been tested for about 1,100 Serbian compound nouns[6], and in Savary et al. (2007) it is used to describe samples of inflectionally non-compositional and irregular compounds in French, Polish and Serbian. In Krstev et al. (2006b) it has been integrated into a platform allowing efficient creation, interconnection and maintenance of heterogeneous linguistic resources, such as simple and compound word lexicons, wordnets, etc. Thus, the annotation of simple constituents within compounds can be automated via real-time access to the underlying DELAF dictionaries.

3.2 Cascaded finite-state approaches

The two-level morphology implemented in the finite-state lexicon compiler, *lexc*, accompanied by the regular-expression compiler, *xfst*, by Beesley and Karttunen (2003), has provided a framework for several approaches to multiword processing.

Lexc

Karttunen et al. (1992) and Karttunen (1993) contain a case study of French compositional and non-compositional compounds. Their morphological description is considered as a typical application for composition of two-level rules.
Firstly, simple words are listed in regular-grammar lexicons, such as the one in example (75). Here a sample noun lexicon contains two words, *démocrate* and *social*, together with their *continuation classes*, *Nmf* and *Adj*. The continuation classes themselves are related to sets of sequences of terminal symbols (*+N, +Adj, +Sg*, etc.) and other continuation classes (*Gender, Number, Masc, Fem*).

(75)
```
Multichar_Symbols +N +Adj +Masc +Fem +Sg +Pl
LEXICON  Root    Nouns ;
LEXICON  Nouns   démocrate Nmf ; social Adj ;
LEXICON  Nm      +N Masc ;
LEXICON  Nmf     +N Gender ;
LEXICON  Adj     +Adj Gender ;
LEXICON  Gender  Masc ; Fem ;
LEXICON  Masc    +Masc Number ;
LEXICON  Fem     +Fem Number ;
LEXICON  Number  +Sg # ; +Pl # ;
```

One possible derivation in this grammar, shown in example (76), allows to obtain the lexical form *social+Adj+Masc+Pl*.

[6] According to a personal communication with Cv. Krstev in May 2008 the number of described Serbian compound lemmas is now 2,822.

(76) *Root* ⟶ *Nouns* ⟶ *social Adj* ⟶ *social +Adj Gender* ⟶
social +Adj Masc ⟶ *social +Adj +Masc Number* ⟶
social +Adj +Masc +Pl #

Further, the lexicon may be composed with a set of lexical alternation rules such as: 'an *l* is replaced by a *u* if it appears after an *a* and before category and gender labels, except *+Adj+Fem*'. This rule allows e.g. to transform the lexical entry *social+Adj+Masc+Pl* into an intermediate form *sociau+Adj+Masc+Pl*. Another possible rule is: 'the *+Pl* label is replaced by an *x* if it appears after an *a* or an *e*, followed by a *u*, followed by category and gender labels, except *+Adj+Fem*'. This second rule applied to *sociau+Adj+Masc+Pl* yields *sociau+Adj+Mascx*. Finally, two other rules 'delete *+Adj*' and 'delete *+Masc*' allow to obtain the surface form *sociaux*. All such alternation rules may be composed into one transducer which maps any lexical form with the corresponding surface form as in example (77).

(77)

démocrate+N+Fem+Sg	*démocrate*
démocrate+N+Fem+Pl	*démocrates*
démocrate+N+Masc+Sg	*démocrate*
démocrate+N+Masc+Pl	*démocrates*
social+N+Fem+Sg	*sociale*
social+N+Fem+Pl	*sociales*
social+N+Masc+Sg	*social*
social+N+Masc+Pl	*sociaux*

Compounds may be treated in a similar lexicalized way, provided that their simple constituents have already been described. Consider the sample lexicon (78), which completes the one in example (75). Each compound appears with its continuation class (here: *AN2mf*). One possible derivation in this lexicon would be as in example (79). The resulting sequence *social-démocrate(0:ˆan)+N+Masc+Pl*, where zero represents an empty string, describes a relation between the first (lexical) and the second (intermediate) form form in (80).

(78)
Multichar_Symbols ˆn ˆan
LEXICON Nouns social-démocrate AN2mf ;
LEXICON AN2mf 0:ˆan Nm ; 0:ˆan Nf ;

(79) *Root* ⟶ *Nouns* ⟶ *social-démocrate AN2mf* ⟶
social-démocrate 0:ˆan Nm ⟶
social-démocrate 0:ˆan +N Masc ⟶
social-démocrate 0:ˆan +N +Masc Number ⟶
social-démocrate 0:ˆan +N +Masc +Pl #

After having added compounds to the lexicon we also enlarge the alternation system by adding new rules allowing feature insertions or propagations

from the whole compound to the individual constituents if the *ˆan* occurs. Examples of such rules are: (i) 'insert *+Adj* at the end of the first constituent (after this operation the *ˆan* marker disappears)', (ii) 'recopy the gender of the whole sequence at the end of the first constituent', (iii) 'recopy the number of the whole sequence at the end of the first constituent'. By applying such rules to the second form in (80) we obtain the third form, which is further transformed into the fourth (surface) form. Similarly, the three other possible derivations within lexicon (78), composed with alternation rules, complete the inflectional paradigm of the compound by describing the surface forms *social-démocrate*, *sociale-démocrate*, and *sociales-démocrates*.

(80)
social-démocrate+N+Masc+Pl
↓
social-démocrateˆan+N+Masc+Pl
↓
social+Adj+Masc+Pl-démocrate+N+Masc+Pl
↓
sociaux-démocrates

The underlying formalism, a lexical transducer, is a mathematically well defined and elegant tool, which allows the whole cascade of rules to be performed in one processing step only. Moreover, the bi-directionality of the transducer allows it to perform both the morphological analysis and generation, i.e. assigning the last (surface) form to the first (lexical) form in (80), and conversely.

Exocentric compounds and inflectional irregularities, as in examples (18)-(26), may be expressed by attributing appropriate continuation classes to compounds in the lexicon, and by designing adequate alternation rules assigned to these classes. For instance, example (23) can be added to the above lexicon via the entries in (81) and two alternation rules: (i) 'insert *+Adj+Masc* at the end of the first constituent if the *ˆamnff* marker appears; after this operation the marker disappears', (ii) 'recopy the number of the whole sequence at the end of the first constituent if the *ˆamnff* marker appears'. The two possible lexicon derivations in this lexicon are (82) and (83). The former relates the first and the second form in (84), the application of the alternation rules above yields the third form in the same example, and the application of alternation rules for simple words results in the fourth form. Derivation (83) introduces no compound-oriented morphological annotation (such as *ˆan* or *ˆamnff*), thus only the alternation rules for the final constituent apply, yielding the second form in (85).

(81)
```
Multichar_Symbols  ˆamnff
LEXICON    Nouns     grand-mère  AmN2f  ;
LEXICON    AmN2f     0:ˆamnff  Nf  ;  Nf  ;
LEXICON    Nf        +N  Fem  ;
```

(82) *grand-mère 0:ˆamnff +N +Fem +Pl #*

(83) *grand-mère +N +Fem +Pl #*

(84)
grand-mère+N+Fem+Pl
↓
grand-mèreˆamnff+N+Fem+Pl
↓
grand+Adj+Masc+Pl-mère+N+Fem+Pl
↓
grands-mères

(85)
grand-mère+N+Fem+Pl
↓
grands-mères

One drawback in this approach is that the cascade of morphophonological and morphosyntactic alternation rules is hard to maintain. Addition, deletion or order change of rules are error-prone as they may modify the conditions in which the previous rule cascade operated. Moreover, it is unclear how one may lexically represent compound lemmas, whose components are not lemmas themselves. For instance, if we wish to attach the first to the last form in (86) then we obtain the second intermediate form in which *vive* is not a lemma. Thus, the simple-word lexicon, which knows how to handle the lexical form *vif+Adj+Fem+Pl* might not cope with *vive+Adj+Fem+Pl* correctly. One solution to this problem is choosing a compound lemma in which all constituents are lemmas themselves, i.e. *mémoire-vif+N+Fem+Pl*, however this is inconvenient for human readers. Another solution is to annotate the constituents in the lexical form with their own lemmas, pretty much as in the DELAC dictionaries (see sections 3.1 and 3.1).

(86)
mémoire vive+N+Fem+Pl
↓
mémoire+N+Fem+Pl vive+Adj+Fem+Pl
↓
mémoires vives

Note that the continuation classes attributed to compound lexical entries, such as *AN2mf*, play a similar role in this formalism as the inflection codes for compounds do in section 3.1. They allow, within each compound lexical form, to trigger the adequate operators producing the desired inflected forms.
The *lexc* formalism includes *flag diacritics* allowing to perform various operations on user-defined variables within the lexicon or alternation rules. Thus, a variable may be set to a particular value, negated or unset, its value may be verified or unified. That allows to express long-distance morphological constraints in order to handle non-concatenative morphosyntactic phenomena. Supposedly, that mechanism can be used for an efficient description of

compound inflection, however few examples are given on this subject in the reference bibliography.

IDAREX

IDAREX Breidt et al. (1996) uses an additional regular expression layer over *lexc* for the description of German multi-word expressions (MWEs), in particular verbal ones, and their variation. Inflected forms are represented by regular expressions which may refer either to base forms or to surface forms of simple components. Morphological features of each component may be restrained. Optional components and insertions may be indicated. Syntactic transformations may be listed within one paradigm. For instance, in the following expression:

(87) *[:den (:schönen) :Schein (:zu) wahren |*
wahren Vfin: (ADV NPnom) ADV* :den (:schönen) :Schein]*

the first line accounts for the infinitive expression *den (schönen) Schein (zu) wahren* ('keep up appearances'), in which the verb *wahren* may take any inflected form, while all other components are limited to their literal forms appearing after the ':' character. The second line describes variants of the same expression in which the verb comes first and is limited to any of its finite forms (*Vfin:*), and adverbs and personal pronouns may be inserted between the verb and the rest of the components, as in *dabei wahrt er immer den Schein*.

Numerical variables *$1*, *$2*, etc. may be assigned to the components of the base form, which allows to generically express omissions, duplications and order changes of components. For instance the following macro can be used instead of numerous complex rules such as (87):

(88) *[$2 Vfin: (ADV* NPron) ADV* $1*
| $1 (:zu) $2 V:]

It expresses the fact that many *Verb Object* idioms in German, such as *den (schönen) Schein wahren* or *die Ohren spitzen* ('prick up one's ears'), may appear either in a finite or an infinite form, with optional adverbial and pronominal insertion. Variables *$1* and *$2* get instantiated to the verb (e.g. *wahren* or *spitzen*) and to the object (e.g. *den (schönen) Schein* or *die Ohren*) of the MWE in question. That instantiation yields a rule similar to (87) for any adequate idiom it is applied to.

Non-compositional and irregular nominal compounds and most of their variants are describable by this formalism. For instance, examples (22) and (46) can be represented either by the specific rules (89) and (90) or by the generic ones (91) and (92).

(89) *[:attorney general N: | attorney N: :general]*

(90) *[:birth date N: | date N: :of :birth]*

(91) *[:$1 $2 N: | $1 N: :$2]*

(92) *[:$1 $2 N: | $2 N: :of :$1]*

The main problem in this formalism seems the fact that no inflectional features may be assigned to MWEs' inflected forms. Thus, one may identify sequences in a corpus but not perform their morphological analysis or generation.

Another important drawback is that even if a unification mechanism has been envisaged it has not been implemented up to our knowledge. Thus, each time a feature agreement takes place in a compound, all the inflected forms have to be enumerated explicitly, which is particularly inefficient for large inflection paradigms, e.g. in Slavic languages. For instance example (39) would need a rule containing several dozens of alternatives if the description were supposed not to admit ungrammatical forms.

Multi-word processor of Turkish

Oflazer et al. (2004) describe a multi-word processor for Turkish, which is a highly-inflective and concatenative language. That tool takes the so-called lexicalized (invariable), semi-lexicalized (morphologically variable) and non-lexicalized (duplication- and contrasting-based) collocations and named entities into account. All those units are contiguous sequences of tokens. Inflectional issues are addressed in all those types of MWUs except the first one.

The MWU processor first runs a text tokenizer, a morphological analyzer and a guesser, all three based on the Xerox finite-state lexicon compiler (Beesley and Karttunen, 2003). Then the MWUs are recognized by a three-stage cascade of Perl rules: first the lexicalized collocations are identified, then the non-lexicalized ones, and finally the semi-lexicalized ones. The rules allow to transform sequences of simple words with their morphological interpretations into compounds with their own morphological features. For instance the sequence in example (94), corresponding to the surface string (93), is transformed into the compound interpretation in example (95)[7].

(93) *uyur uyumaz* (literally, '(he) sleeps (he) does not sleep')

(94) *uyu+Verb+Pos+Aor+A3sg uyu+Verb+Neg+Aor+A3sg*

(95) *uyu+Verb+Pos+^DB+Adverb+AsSoonAs* ('as soon as he sleeps')

Unfortunately, no example of such a Perl rule used in this approach in given in the reference paper but we suppose that those rules are of two types: (i) lexicalized rules, which allow to replace sequences of precise lexical units

[7] The morphological codes used are: *+Pos*: positive polarity, *+Neg: negative polarity*, *+Aor*: aorist aspect *+A3sg*: third person singular, *^DB*: derivation boundary

by compounds containing the same units, if only some morphological constraints are respected for these units; thus, they are roughly equivalent to inflection codes for compounds described in section 3.1, (ii) non-lexicalized rules, which match duplications of any lexical item, sometimes accompanied by extra elements such as contrasting or clitics.
Although this approach uses the same lexicon compiler as in section 3.2, the important difference seems to be the fact that the relation between sequences of tokens and compounds containing those tokens is implemented by Perl rules instead of a lexical transducer, which probably does not allow the reversibility of those rules.

3.3 Relational database approaches

Naturally enough, the behavior of simple words both as individual units, and as members of compounds, may be represented by a relational database.

HABIL

HABIL (Alegria et al., 2004) is a multi-word expression processor for Basque. It deals with both contiguous and split MWEs, either totally fixed and opaque, or decomposable. Lexicalized compounds are included in this set. For each MWE it admits reorderings of its simple components, and checks their inflectional constraints. It also generates morphosyntactic interpretations for the MWEs so that ambiguous structures can get multiple representations.
It is based on a relational database containing morphological descriptions of 80,000 simple words and 2,270 MWEs. On the one hand, each MWE is related to its simple components, each of which is described in particular with respect to: (i) its position and the inflected form that it takes in the MWE's lemma, (ii) its inflectional paradigm, (iii) if it is or not the headword of the MWE. Each component, via its lemma and its homograph number, is related to its inflectional paradigm. For instance table 1 describes the three components of the MWE *begi bistan egon* ('to be evident'), the second of which is the inflected form *bistan* of the lemma *bista* and the homograph number 1. That description resembles the morphological tags that simple constituents get in the Multiflex DELAC entry (see section 3.1).
On the other hand a MWE is related to all its possible *surface realizations* (i.e. inflected forms), where the order of the components and possible insertions of extra elements are described, together with inflectional restrictions imposed on all possibly inflected components. For instance, the upper part of table 2 describes three unambiguous[8] surface realizations of the MWE, whose lemma is *begi bistan egon*. In the last realization the order of components (*3?12*) is different than in the lemma (the 3rd constituent comes first) and one insertion between components 3 and 1 is allowed. Its first compo-

[8] Their occurrences in a corpus may never be analyzed as sequences of "free" words.

Entry	Homograph ID	Component position	Component inflected form	Headword	Component lemma	Component homograph ID
begi bistan egon	0	1	*begi*	*no*	*begi*	2
begi bistan egon	0	2	*bistan*	*no*	*bista*	1
begi bistan egon	0	3	*egon*	*yes*	*egon*	1

TABLE 1 *HABIL* table describing the components of the MWE *begi bistan egon*

Entry	Homograph ID	Order & contiguousness	Unambiguous	Inflection restrictions
begi bistan egon	0	123	*yes*	component 1 : ((CAS=ABS) and (DEF=-)) or ((CAS=GEN) and (NUM=PL)) component 2 : *fixed* component 3 : *any form*
begi bistan egon	0	312	*yes*	component 1 : ((CAS=ABS) and (DEF=-)) or ((CAS=GEN) and (NUM=PL)) component 2 : *fixed* component 3 : *any form*
begi bistan egon	0	3?12	*yes*	component 1 : ((CAS=ABS) and (DEF=-)) component 2 : *fixed* component 3 : *any form*
toile d'araignée	0	123	*no*	component 1 : *any form* component 2 : *fixed* component 3 : *fixed*
toile d'araignée	0	123	*no*	component 1 : NUM=PL component 2 : *fixed* component 3 : NUM=PL

TABLE 2 *HABIL* table describing the surface realizations of the Basque MWE *begi bistan egon* and of the French MWU *toile d'araignée*

nent (*begi*) is inflected into absolutive non-definite form, while its second one (*bistan*) remains uninflected, and its third one (*egon*) may be inflected to any of its existing forms. These constraints may correspond for instance to the following corpus occurrence: *ez dago horren begi bistan* ('it is so evident').
This approach allows for an exhaustive description of inflected forms of a MWE, together with some of its variants resulting from omissions, duplications, and order changes of constituents. Thus, examples like (23) through (29), (38) through (40), and (47) through (49) seem describable in this model. For instance, the lower part of table 2 shows a possible description of example (24), in which two rules are necessary to express the plural variant, but the corresponding lemma is unique (unlike examples (56)-(58)).
The formalism also allows for non-abstract base forms (cf. examples (48) and (49)), as well as for insertions, however the inserted elements may not be specified, which is needed in examples like (37), (45) and (46). It is unclear how the inflection features are determined for exocentric compounds (examples (18) through (22)). One drawback, for languages with large inflectional paradigms (see section 2.5), is that if agreement constraints occur within a MWE then each of its inflected forms needs a separate entry in the database (e.g. several dozens of entries for most compound nouns in Slavic languages). An additional unification mechanism could solve this inconvenience. Another disadvantage results from the fact that separators are not considered as components, thus it seems impossible to account for their insertions, deletions or replacements (cf. examples (8) and (45)).

3.4 Unification grammar approaches

As mentioned above, one of the reasons why MWUs are placed on the frontier between morphology and syntax, is that dependencies of different kinds occur among MWU constituents, such as agreement and government rules. Thus, naturally enough, the description of MWUs has been addressed within several frameworks based on unification grammars.

LinGO project

In Sag et al. (2002), Copestake et al. (2002) and Villavicencio et al. (2004), a large project of syntactic and semantic description of English multi-word expressions is discussed. Their syntax is described with typed feature structure (TFS) grammars implemented within a constraint-based HPSG formalism, while their semantics is addressed within the Minimal Recursion Semantics formalism. General grammar rules describe general language phenomena, while lexicalized rules, implemented as a relational database, are introduced to account for idiosyncrasy. This database approach differs from the one in section 3.3 in that the database model is not supposed to reflect the language model, but only to provide an internal representation of the TFS rules.

The MWEs are divided into several classes with respect to their semantic compositionality and their syntactic variability.
Fixed expressions (e.g. *by and large, every which way, ad hoc*) are seen as 'words with spaces' as they defy conventions of grammar and admit no morphological or lexical variability. For instance, example (96) describes *ad hoc* as a simple concatenation of two tokens, functioning as an intransitive adjective (*intr_adj_l*) and allowing no syntactic variation.

(96) *ad_hoc_1 :=* *intr_adj_l &*
[STEM <"ad","hoc">,
SEMANTICS [KEY ad-hoc_rel]]

Semi-fixed expressions admit inflection, or selection of determiners or reflexive forms. They are further divided into three classes: (i) non-decomposable idioms (e.g. *kick the bucket, wet oneself, shoot the breeze*) characterized by the semantic non-decomposability, lack of syntactic and lexical variation (**kick the great bucket in the sky*[9], **the breeze was shot*), but allowing inflection (*kicked the bucket*) and variation in reflexive form (*wet myself*), (ii) compound nominals (e.g. *car park, attorney general*) which do not admit syntactic variants but inflect in number, (iii) proper names (e.g. *the San Francisco 49ers*), which are highly idiosyncratic (**the Oakland 49ers*), may require a definite article (*the, those*) and allow for omission and insertion variants (*the 49ers, the league-leading San Francisco 49ers*).
Syntactically-flexible expressions split into three classes: (i) verb-particle constructions are either semantically idiosyncratic (*brush up on*) or compositional (*call up, eat up, fall off*), may admit insertions (*call Kim up* but **fall a truck off*), and usually show syntactic idiosyncrasy (*call/ring/telephone, call up/ring up* vs. **telephone up*), (ii) decomposable idioms (*spill the beans*) reveal a quite compositional semantics (*spill* $\approx$ *reveal, the beans* $\approx$ *a secret*) and are syntactically partially flexible but this flexibility is unpredictable, (iii) light verbs (*make* in *make a mistake*) are used in an unpredictable way (**do a mistake*), their sense is bleached but their complement noun phrases are used in the normal sense, and the whole constructions admit full syntactic variability (*a mistake was made, make a big mistake*). Syntactically-flexible expressions are treated by inheritance hierarchies and lexical selection.
Institutionalized phrases (*traffic light*) are semantically and syntactically compositional but statistically idiosyncratic (**intersection regulator*). The technique used for their treatment is not specified in the reference papers.
The inflection of MWEs is clearly not a major issue in the reference papers. It is possible to point out the internal component that allows inflection. For

[9] The authors don't mention the exceptional variations such as *kick the proverbial bucket* attested in corpora.

instance, example (97) describes the nominal compound *part of speech* in which only the first component inflects.

(97) *part_of_speech_1 :=* *intr_noun_l &*
[STEM <"part","of","speech">,
INFL-POS "1",
SEMANTICS [KEY part_of_speech_rel]]

As the LinGO rules are expressed in what seems a full-fledged grammatical formalism, component agreements should be describable (e.g. *he dined and wined Susan*), however no example of such a rule is given in the reference papers. Embedding, inheritance and optionality mechanisms allow modular description, pointing out headwords, and describing omissions like in example (38).

It is unclear if it is possible to express limiting the inflection to some forms only (as in examples (28) and (29)), inflection variants (as in examples (21) through (24)), inversions, insertions and duplications (as (37), (39), (40), (45) and (46)), as well as the exocentricity of compounds (as (18) through (22)).

FASTR

FASTR (Jacquemin, 2001) is a shallow parser dedicated to the recognition, normalization and acquisition of compound terms, developed within a unification-based framework. Its design was based on an in-depth study of inflectional, syntactic and semantic variation of terms in English and French in various specialized domains. FASTR's input is a corpus and an initial set of controlled complex terms that are analyzed morphologically and transformed into feature structure rules. Its output is a set of links between initial terms and occurrences of these terms and their variants in the corpus.

For instance rule (98) is a 'flat' representation of the feature structure (99) resulting from the term *umbilical artery*. The rule is fully lexicalized: the lemmas of both components are explicitly indicated. Their *inflection* features are inspired by the inflection codes for simple words in the DELA methodology (cf. section 3.1). The *lexicalization* feature allows to link the rule to the lexical item *arter* called the lexical anchor. The *agreement* features allow to express how morphological features are constrained or propagated : here, the number of the lexical anchor (N_3) is propagated to the whole compound (N_1). Thus, the rule matches both the singular and the plural occurrences of the term: *umbilical artery* and *umbilical arteries*.

(98) *Rule* $N_1 \rightarrow A_2\ N_3$*:*
$\langle N_1$ *lexicalization*$\rangle \doteq N_3$
$\langle A_2$ *lemma*$\rangle \doteq$ *'umbilical'*; $\langle A_2$ *inflection*$\rangle \doteq$ *1*
$\langle N_3$ *lemma*$\rangle \doteq$ *'arter'*; $\langle N_3$ *inflection*$\rangle \doteq$ *2*
$\langle N_1$ *agreement*$\rangle \doteq \langle N_3$ *agreement*$\rangle$.

(99)

$$\begin{bmatrix} \textit{arity}: & 2 & & & \\ & & \textit{lexicalization}: & \boxed{1} & \begin{bmatrix} \textit{lemma}: & \text{`}\textit{arter}\text{'} \\ \textit{cat}: & \text{`}N\text{'} \\ \textit{inflection}: & 2 \\ \textit{agreement}: & \boxed{2} \end{bmatrix} \\ 0: & & \textit{cat}: & \text{`}N\text{'} & \\ & & \textit{agreement}: & \boxed{2} & \\ 1: & \begin{bmatrix} \textit{lemma}: & \text{`}\textit{umbilical}\text{'} \\ \textit{cat}: & \text{`}A\text{'} \\ \textit{inflection}: & 1 \end{bmatrix} & & & \\ 2: & \boxed{1} & & & \end{bmatrix}$$

Terminological variants are expressed in FASTR via transformations represented by *metarules* (a concept introduced in a number of unification-based formalisms in order to reduce the grammar size). For instance, metarule (100), when unified with rule (98), produces the new rule (101) which matches coordination variants such as *umbilical or carotid artery*.

(100) *Metarule Coord*($N_1 \rightarrow A_2\ N_3$) $\equiv N_1 \rightarrow A_2\ C_4\ A_5\ N_3$:.

(101) *Rule* $N_1 \rightarrow A_2\ C_4\ A_5\ N_3$:
$\langle N_1$ *lexicalization*$\rangle \doteq N_3$
$\langle A_2$ *lemma*$\rangle \doteq$ *'umbilical'*; $\langle A_2$ *inflection*$\rangle \doteq$ *1*
$\langle N_3$ *lemma*$\rangle \doteq$ *'arter'*; $\langle N_3$ *inflection*$\rangle \doteq$ *2*
$\langle N_1$ *agreement*$\rangle \doteq \langle N_3$ *agreement*$\rangle$.

Metarules also allow to express derivational variants, provided that the derivational morphology of the simple components is described. For instance, the sequence *tension artérielle* in example (41) may be expressed by the compound term rule (104), while the variant *tension des artères* is obtained by unifying metarule (105) with rule (104) and with the word descriptions (102) and (103)[10]. The key element here is the first constraint of rule (105) imposing that the second noun of the variant (here: *artères*) has the same root as the adjective of the base term (here: *artérielle*).

(102) *Word 'artère':*
$\langle$*cat*$\rangle \doteq$ *'N'*; $\langle$*secondary root*$\rangle \doteq$ *'artér'*; $\langle$*inflection*$\rangle \doteq$ *21*.

(103) *Word 'artériel':*
$\langle$*cat*$\rangle \doteq$ *'A'*; $\langle$*inflection*$\rangle \doteq$ *2*; $\langle$*root cat*$\rangle \doteq$ *'N'*;
$\langle$*root lemma*$\rangle \doteq$ *'artère'*; $\langle$*history*$\rangle \doteq$ *'?ielle'*.

(104) *Rule* $N_1 \rightarrow N_2\ A_3$:
$\langle N_1$ *lexicalization*$\rangle \doteq N_2$

[10]The '?' sign in the derivation suffix '?ielle' refers to the secondary lemma *artér* of the word whose the main lemma is *artèr*

$\langle N_2\ lemma\rangle \doteq$ *'tension'*; $\langle N_2\ inflection\rangle \doteq 1$
$\langle A_3\ lemma\rangle \doteq$ *'artériel'*; $\langle A_3\ inflection\rangle \doteq 2$
$\langle N_1\ agreement\rangle \doteq \langle N_2\ agreement\rangle \doteq \langle A_3\ agreement\rangle$.

(105) *Metarule AdjToNoun*$(N_1 \rightarrow N_2\ A_3) \equiv N_1 \rightarrow N_2\ D_4\ N_5$*:*
$\langle A_3\ root\rangle \doteq \langle N_5\ root\rangle$
$\langle D_4\ lemma\rangle \doteq$ *'de'*; $\langle D_4\ inflection\rangle \doteq 1$
$\langle D_4\ agreement\ number\rangle \doteq \langle N_5\ agreement\ number\rangle$

Since in exocentric compounds, such as (18) through (22), the feature propagation cannot be performed, the morphology of the resulting compound must be indicated explicitly, as in rule (106).

(106) *Rule* $N_1 \rightarrow V_2\ N_3$*:*
$\langle V_2\ lemma\rangle \doteq$ *'perce'*; $\langle V_2\ inflection\rangle \doteq 1$;
$\langle V_2\ agreement\ tense\rangle \doteq$ *'present'*; $\langle V_2\ agreement\ person\rangle \doteq 3$
$\langle V_2\ agreement\ mood\rangle \doteq$ *'indicative'*
$\langle N_3\ lemma\rangle \doteq$ *'neige'*; $\langle N_3\ inflection\rangle \doteq 1$;
$\langle N_3\ agreement\ number\rangle \doteq$ *'singular'*
$\langle N_1\ agreement\ gender\rangle \doteq$ *'masculine'*
$\langle N_1\ agreement\ number\rangle \doteq$ *'singular'* | *'plural'*.

Compounds admitting variants of inflected forms, as in examples (21) through (24), may also be described via metarules, which however need to be lexicalized in order to avoid spurious variants for regular constructions. For instance, rule (107) matching *attorney general* and *attorney generals*, when unified with metarule (108), matches the plural variant *attorneys general*. The 5th constraint in rule (107) does not allow to interpret *attorneys general* as a singular form.

(107) *Rule* $N_1 \rightarrow N_2\ N_3$*:*
$\langle N_1\ lexicalization\rangle \doteq N_2$
$\langle N_2\ lemma\rangle \doteq$ *'attorney'*; $\langle N_3\ lemma\rangle \doteq$ *'general'*
$\langle N_2\ inflection\rangle \doteq \langle N_3\ inflection\rangle \doteq 1$
$\langle N_2\ agreement\ number\rangle \doteq$ *'singular'*;
$\langle N_1\ agreement\rangle \doteq \langle N_3\ agreement\rangle$.

(108) *Metarule DoublePlural*$(N_1 \rightarrow N_2\ N_3) \equiv N_1 \rightarrow N_4\ N_5$*:*
$\langle N_2\ lemma\rangle \doteq \langle N_4\ lemma\rangle \doteq$ *'attorney'*
$\langle N_3\ lemma\rangle \doteq \langle N_5\ lemma\rangle \doteq$ *'general'*
$\langle N_3\ agreement\ number\rangle \doteq \langle N_4\ agreement\ number\rangle \doteq$ *'plural'*
$\langle N_5\ agreement\ number\rangle \doteq$ *'singular'*.

Variation schemes which appear systematically up to some exceptions may be expressed by general (i.e. non-lexicalized) metarules accompanied by negative metarules. For instance, the sequence *bezwzględna większość* in example (39) may be represented by rule (109), while its variant *większość bezwzględna* results from unifying this rule with metarule (110). However, a

number of *Adjective Noun* compounds in Polish, such as *dobre imię* ('a good reputation'), do not admit inversion of their constituents. Such exceptions may be described by negative metarules. For instance, the negative metarule (111), when unified with a rule describing *dobre imię*, analogous to rule (109), matches the invalid variant **imię dobre*. During corpus processing FASTR rejects all sequences that have been matched by negative metarules, thus **imię dobre* will not be admitted as a variant of *dobre imię*.

(109) *Rule* $N_1 \rightarrow A_2\ N_3$*:*
$\langle N_1$ *lexicalization*$\rangle \doteq N_3$
$\langle A_2$ *lemma*$\rangle \doteq$ *'bezwzględny'*; $\langle A_2$ *inflection*$\rangle \doteq$ *1*
$\langle N_3$ *lemma*$\rangle \doteq$ *'większość'*; $\langle N_3$ *inflection*$\rangle \doteq$ *45*
$\langle N_1$ *agreement*$\rangle \doteq \langle A_2$ *agreement*$\rangle \doteq \langle N_3$ *agreement*$\rangle$.

(110) *Metarule InvPlural(*$N_1 \rightarrow A_2\ N_3$*)* $\equiv N_1 \rightarrow N_3\ A_2$*:*.

(111) *Metarule NInvPlural(*$N_1 \rightarrow A_2\ N_3$*)* $\equiv N_1 \rightarrow N_3\ A_2$*:*
$\langle A_2$ *lemma*$\rangle \doteq$ *'dobre'*; $\langle N_3$ *lemma*$\rangle \doteq$ *'imię'*.

In conclusion, FASTR is a highly specialized corpus processor for the recognition of inflectional, syntactic and semantic variants of complex terms. It is based on terminological vocabulary transformed in an automated way into a set of linguistically annotated rules, and can be tuned to each new application and sublanguage. Its unification-based formalism offers a great expressive power for the description of various linguistic constraints.
The main drawback of this approach is that it is not very modular. The modeling of term variation relies on a large set of interfering rules, which is not always easy to maintain. Firstly, a modification of one rule may influent the correctness of other rules[11]. Secondly, checking if a particular term is correctly described requires, unlike in the DELA methodology (cf. section 3.1), controlling several rules simultaneously which is not optimal for a human lexicographer.

4 Comparative study

Tables 3 and 4 present a comparative summary of the approaches presented in section 3. The features appearing in the first column correspond to the linguistic properties of MWUs discussed in section 2. The meaning of a '✓' character, a '×' character, and a '?' character is, respectively, that the corresponding approach accounts for the given property, it does not account for the property, or it is unclear if it does. In particular, we suppose that:

- *Separators* are allowed to have a status of MWU's constituents, if examples like (3) through (5) can be described, and if the sequences in example (8)

[11] The same criticism was leveled against the cascaded finite-state morphology models (cf. section 3.2), however in FASTR the degree of the dependency between rules is much lower.

can be distinguished but attached to the same lemma.

- *Squeezed compounds* are those in which a boundary between constituents is not marked by a separator, as in examples (11) through (14). They should be morphologically described as sequences of possibly inflected simple words.
- *Exocentric compounds* are describable, if examples like (18) through (22) can be properly treated.
- *Irregular agreement* is properly treated if all forms in examples like (23) through (26), as well as (2) and (16), can be described, and if no introduction of artificial lemmas is needed (cf. (56) for a counterexample).
- *Defective paradigms* can be described if examples like (28) and (29) don't suffer from overgeneralization, i.e. their inexistent singular forms are not accepted.
- *Insertions and omissions* are accounted for, if variants containing extra elements, as *en* in example (37), can be attached to a lemma which does not contain this element, and if variants missing some constituent, can be attached to the lemma in which this constituent appears, as in (38).
- *Order change* is taken into account, if variants like (39) can be attached to the same lemma.
- *F*orms resulting from component *duplication* should be attached to a lemma where this component is not duplicated, as in example (40).
- *Derivational and semantic variants* should be related to their base forms containing no derived form and no semantic replacement, as in (41) and (42).
- *Abbreviations* should be attached to their full forms, as in (43) and (44).
- *Unification* is necessary for a compact representation of huge inflection paradigms of MWUs, especially those in which agreement rules apply within constituents (cf. example (15) and section 2.5).
- The lemma of a MWU is *non-abstract*, if it is a linguistically correct form (cf. examples (48) and (49)).
- *Non-contiguous MWUs* are treated, if extra elements, not belonging to an inflected form, are admitted within this form in a corpus, as in example (50).
- The morphological description of MWUs is *non-redundant* if there is a unique representation of the inflectional behavior of simple words appearing in MWUs (cf. section 3.1 for a counterexample).
- *Inflectional analysis and generation* are two computational applications for which a MWU description module should be accessible.
- An *automated MWU lexicon creation* is a facility of a computational platform allowing to avoid as much manual lexicographic work as possible. It may rely for instance on exploitation of the existing resources for simple words in order to annotate the components of MWUs.

- The *sense computation* is the capacity of representing the meaning of a MWU, if possible, as a function of the meanings of its components (cf. section 3.4).
- The *formal tool* is a theoretical framework used either for the description of MWUs, or for their internal representation and treatment.
- The *number of MWUs described* refers to the MWUs' base forms, and not their inflected forms.
- The *language* indicated is the one concerned by the experiments described in the bibliography.

The data presented in tables 3 and 4 confirm the importance of compositional phenomena in natural languages. Different NLP schools have been studying these phenomena to a varying extent, and those presented here propose lexicalized approaches, i.e. multi-word units are explicitly listed and their linguistic behavior is described either by explicit shared paradigms (e.g. inflectional codes in the DELA school), or by lexicalized grammars in which separate rules may interfere (e.g. alternation rules in *lexc*, or rules and metarules in *FASTR*). One interesting type of MWUs, duplications (cf. section 3.2), has been treated by non-lexicalized patterns.

The results presented are quantitatively very different. Some approaches rely only on samples of less than several hundred entries, some others judge one or two thousand entries as sufficiently representative, while the remaining ones have achieved a large-scale description of tens of thousands of MWUs. In particular, most features appearing in table 4 for *lexc* imply the pertinence of this approach to the morphological treatment of MWUs, however, they need an experimental confirmation in real-size MWU lexicons.

The linguistic properties discussed in section 2 are only partly addressed in the references papers. The appreciation of these phenomena is not necessarily better with a growing number of the entries described.

Some particularly discriminating features are:

- Separators, whose role in MWUs is underestimated by the majority of the approaches.
- Some idiosyncratic aspects of the inflection of MWUs (exocentricity and agreement irregularities), which are not addressed by some approaches, although they belong to the fundamental properties of these units.
- Defective paradigms whose importance has been identified by virtually all approaches.
- Derivational and semantic variants of MWUs, which are explicitly treated only by *FASTR* (we suppose though *lexc*'s and *LinGO*'s pertinence for derivational and semantic variants, respectively).
- Abbreviations which are explicitly addressed by none of the approaches studied, expect possibly *NooJ*, which however may handle them by a rather

brute-force method only. Integrating abbreviations into the morphology of simple words (e.g. by saying that *phys* is a possible realization of *physical*) is a possible indication that they have been taken into account.

- Non-contiguous MWUs (or multi-word expressions) which are increasingly taken into account.
- A non-redundant component description admitted by all but one approach. Thus, the morphology of simple words is described first, and then MWUs are seen as combinations of particular forms of simple words.
- The fact of allowing for morphological analysis and generation of MWUs. Most approaches take both these applications into account.
- The automated lexicon creation. Surprisingly enough, it has been proposed by only two approaches presented, although it seems inevitable if lexicalized approaches to compounding are to efficiently reach large-scale dimensions.
- The sense computation of multi-word units which seems distant from the morphological description of those units. Although judged as non-operational by Downing (1977), it has been addressed by several semantic studies (e.g. Fabre and Sébillot, 1996). However, up to our knowledge, no existing approach focuses equally on both morphosyntactic and semantic phenomena in MWUs within a unique framework. *LinGO*, the only approach presented here which aims at sense computation, leaves many morphological questions unanswered.

Our comparative study does not allow for a clear-cut ranking of the approaches presented. None of them takes all of the linguistic properties and applications discussed into account, or it is unclear if they do. Those that seem particularly attentive to morphosyntactic flexibility of MWUs are *Multiflex*, *lexc* and *FASTR*, as well as *HABIL*, which needs an additional unification mechanism. However few hints exist with respect to how these four models could be extended to cover a full range of non-contiguous multi-word units.
An interesting question, that was hardly addressed in the reference papers, is how to represent MWUs whose individual components may enter into dependency relations with external elements. Intuitively, *LinGO* and *FASTR* should be most adapted to such cases due to the use of feature structures.
Note that our comparative study does not show how the different approaches deal with deciding what should and what shouldn't be considered as a multi-word unit. As mentioned in section 1, avoiding these considerations is deliberate due to the existence of numerous and highly controversial views. Moreover, only some of the approaches presented here (the DELA approaches and Sag et al., 2002) refer to in-depth linguistic theories proposing defining criteria of compounds and other MWUs, and these could hardly be resumed by a checklist. Nevertheless, an *operational* definition of a MWU is clearly a lin-

guistically crucial question, in particular in the context of lexical approaches, for which the first step is to decide if a unit should or shouldn't be included in the lexicon.

5 Recommendations and perspectives

Our comparative study shows a large variety of formal models and computational tools proposed to account for morphological, morphosyntactic and semantic variation of multi-word units in corpora. At the same time, the ongoing work on standards, such as Calzolari et al. (2002) and ISO/TC 37/SC 4 (2007), aims at establishing a common exchange framework for lexical resources. Multi-word units and expressions are increasingly accounted for in these proposals, however their morphology is not as extensively handled as that of simple words. We hope that the following recommendations resulting from our study may contribute to this discussion:

- **A variety of natural languages should be taken into account during elaboration of standards**. The predominating position of the English language has prevented the NLP research from a full appreciation of the importance of morphological phenomena in multi-word units. Taking into account lesser studied, often inflectionally rich, languages, such as Slavic or concatenative languages, should lead to more universal models, platforms and standards.
- For instance, the study of these languages argues for the **necessity of a unification mechanism** for a compact description of agreement rules between components, as well as of huge inflectional paradigms (cf. example (70)).
- If we wish to provide a reusable and universal morphological resource of MWUs, it is important to **take at least the two most general linguistic applications into account: the morphological analysis and generation**. In particular, it should be possible not only to identify a MWU in a corpus but also to annotate it with morphological features necessary for further processing stages. Approaches like *IDAREX*, that don't allow for annotation, seem satisfactory only for a limited number of applications (e.g. concordancers).
- On the very basic graphical level, the NLP community is still far from reaching a consensus on what should be considered as an elementary indivisible unit. Morphological analyzers of simple words differ on this point, even with respect to the same natural language. However, defining the graphical frontier between lexical units is necessary, as it influences the way how multi-word units are defined and processed. We think that **the definition of a lexical unit should be flexible, and adaptable to each new language or application**. In particular, it should be possible to **describe squeezed compounds as sequences of simple words**. Conversely,

		French DELAC (1993)	English DELAC (2000)	Greek DELAC (2002)	NooJ DELAC (2005)	*Multiflex* (2005)
1	Separators as constituents	×	×	×	✓	✓
2	Squeezed MWUs	×	×	×	×	✓
3	Exocentric MWUs	✓	✓	×	✓	✓
4	Irregular agreement	×	✓	×	✓	✓
5	Defective paradigms	✓	✓	✓	✓	✓
6	Insertions and omissions	×	×	×	✓	✓
7	Order change	×	×	×	×	✓
8	Duplications	×	×	×	×	✓
9	Derivational variants	×	×	×	×	×
10	Semantic variants	×	×	×	×	×
11	Abbreviations	×	×	×	✓	×
12	Unification	×	✓	×	×	✓
13	Non-abstract lemmas	×	✓	×	✓	✓
14	Non-contiguous MWUs	×	×	×	×	×
15	Non-redundancy	✓	✓	✓	×	✓
16	Infl. analysis	✓	✓	✓	✓	✓
17	Infl. generation	✓	✓	✓	✓	✓
18	Automated MWU lexicon creation	×	×	×	×	✓
19	Sense computation	×	×	×	×	×
20	Formal tool	text filters, FSTs	sublists, FSTs	restriction filters, FSTs	cut-and-paste rules, FSTs	graphs, FSTs
21	Number of MWUs described	126,000	60,000	27,000	?	2,822
22	Language	French	English	Greek	English	Serbian

TABLE 3 Comparative features of tools for MWU inflectional description. DELA dictionaries.

		lexc (1992)	*IDAREX* (1996)	Oflazer et al. (2004)	*HABIL* (2004)	*LinGO* (2004)	*FASTR* (2001)
1	Separators as constituents	✓	✓	?	×	×	✓
2	Squeezed MWUs	✓	?	?	×	×	×
3	Exocentric MWUs	✓	✓	?	?	?	✓
4	Irregular agreement	✓	✓	?	✓	?	✓
5	Defective paradigms	✓	✓	?	✓	?	✓
6	Insertions and omissions	✓	✓	?	✓	?	✓
7	Order change	✓	✓	?	✓	?	✓
8	Duplications	✓	✓	✓	✓	?	✓
9	Derivational variants	✓	?	?	?	?	✓
10	Semantic variants	×	×	×	×	?	✓
11	Abbreviations	×	×	×	×	×	×
12	Unification	✓	×	?	×	✓	✓
13	Non-abstract lemmas	?	?	?	✓	?	✓
14	Non-contiguous MWUs	×	✓	?	✓	✓	×
15	Non-redundancy	✓	✓	✓	✓	✓	✓
16	Infl. analysis	✓	×	✓	✓	✓	✓
17	Infl. generation	✓	×	?	✓	✓	×
18	Automated MWU lexicon creation	×	×	×	×	×	✓
19	Sense computation	×	×	×	×	✓	×
20	Formal tool	FSTs	FSTs, regular expressions	FSTs Perl rules	relational database	HPSG rules	feature structures, metarules
21	Number of MWUs described	10	?	1,110	2,270	125	72,000
22	Language	French	German	Turkish	Basque	English	English

TABLE 4 Comparative features of tools for MWU inflectional description. Other approaches.

sequences containing blanks (e.g. *a priori*) **should be describable as indivisible tokens**. Moreover, it should be possible to **view separators, punctuation marks, digits, etc. as full members of MWUs** and allow to describe their absence, presence and variation.

- For an efficient human usage and treatment, non-abstract lemmas of MWUs should be offered to lexicographers. As seen in section 2.5, a lemma of a MWU may contain simple words that are not lemmas themselves. Thus, **avoiding abstract multi-word base forms requires the annotation of simple components with their own base forms and features**, as in the English DELAC, *Multiflex* and *HABIL*.
- **The extensiveness of orthographic, morphological, syntactic and semantic variation calls for a common descriptive framework** in which all those types of variations could be taken into account. Here again, lesser studied languages, such as Turkish, reveal new types of morphosyntactic variants such as duplications.
- In order to express omissions, insertions and order changes, it is necessary to **refer to the position of a single component in a compound**. In the existing approaches that may be done either by numbering lexical items (as in *IDAREX*, *Multiflex*, *HABIL* and *FASTR*), or by regular expressions that identify token frontiers (as in *lexc*).
- Most often, morphological forms that simple words take within MWUs, are subsets of the inflectional paradigms of these words. Thus, it seems most natural to admit a **'two-layer' approach**[12]:
  - Describing the morphology of simple words as individual units.
  - Describing multi-word units as morphologically and syntactically conditioned compositions of simple words and other lexical items, such as separators, digits, etc.

  Approaches, such as *NooJ*, in which this postulate is not assumed, suffer from a too high degree of redundancy in component morphology description.
- Studies on the morphological treatment of simple words have been developed for decades and resulted in a large number of formalisms and tools in various languages. Rather than impose a uniform framework both for simple words and MWUs, it seems reasonable to encourage **modularity and interoperability**. Thus, a morphological module for MWUs should be able to interact with any such module for simple words, provided that some interface constraints have been properly defined and respected.
- In order to reach large-scale dimensions in MWU resources, **tools for automated lexicon enrichment should be integrated** into the descriptive process.

[12] Not to be confused with Koskenniemi's two-level morphology.

- Non-contiguous MWUs, as well as their sense computation, remain a challenge. Studies dedicated to multi-word expressions should focus as much on their morphological constraints as on their semantic complexity.

A substantial effort has been made in ISO/TC 37/SC 4 (2007) in order to propose the Lexical Markup Framework (LMF), an abstract metamodel for computational lexicons. In this normative proposal the so-called core package allows to define senses and definitions of lexical entries. It is accompanied by specialized extensions, three of which are of main interest for us: the morphology extension, the NLP paradigm pattern extension, and the NLP multi-word expression patterns extension. While the morphology of simple words is documented there with numerous examples, in which different languages and varying formalisms are expressed, no examples are given of how inflected forms of MWUs can be expressed in terms of inflected forms of their components. Investigating how different lexical approaches to MWU morphology can be mapped onto to this proposal, and an extension of the norm if necessary, are necessary tasks.

Acknowledgments

The author is grateful to Jean-Yves Antoine, as well as to three anonymous reviewers, for their critical remarks on a previous version of this study.

References

Alegria, Inaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and Treatment of Multiword Expressions in Basque. In *Second ACL Workshop on Multiword Expressions, July 2004*, pages 48–55.

Anscombre, Jean-Claude. 1990. Pourquoi un moulin à vent n'est pas un ventilateur. *Langue Française* 86:103–125.

Baldwin, Timothy and Aline Villavicencio. 2002. Extracting the Unpredicatble: A Case Study on Verb-particles. In *Sixth Conference on Computational Natural Language Learning (CoNLL)-2002*, pages 99–105.

Bauer, Laurie. 1983. *English Word-Formation*. Cambridge University Press.

Beesley, Kenneth R. and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI.

Benveniste, Emile. 1974. *Fondements syntaxiques de la composition nominale. Formes nouvelles de la composition nominale*, pages 145–176. Gallimard, Paris.

Breidt, Elisabeth, Frédérique Segond, and Guiseppe Valetto. 1996. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of COLING-96, Copenhagen*, pages 1036–1040.

Cadiot, Pierre. 1992. A entre deux noms : vers la composition nominale. *Lexique* 11:193–240.

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards Best Practice

for Multiword Expressions in Computational Lexicons. In *LREC'02*, pages 1934–1940.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *LREC 2002*, pages 1941–1947.

Corbin, Danielle. 1992. Hypothèses sur les frontières de la composition nominale. *Cahiers de grammaire* 17:26–55. Université de Toulouse Le Mirail.

Courtois, Blandine and Max Silberztein, eds. 1990. *Les dictionnaires électroniques du français*. Larousse, Langue française, vol. 87.

Downing, Pamela. 1977. On the Creation and Use of English Compound Nouns. *Language* 53(4):810–842.

Fabre, Cecile and Pascale Sébillot. 1996. Interprétation automatique des composés nominaux anglais hors domaine: quelles solutions. In *10e Congrès Reconnaissance des Formes et Inelligence Artificielle (RFIA'96), Rennes*, pages 71–79.

Grefenstette, Gregory and Pasi Tapanainen. 1994. What is a word, what is a sentence? problems of tokenization. In *3rd International Conference on Computational Lexicography*, pages 79–87.

Gross, Gaston. 1988. Degré de figement des noms composés. *Langages* 90:57–71. Paris : Larousse.

Gross, Gaston. 1990. Définition des noms composés dans un lexique-grammaire. *Langue Française* 87:84–90.

Gross, Gaston. 1996. *Les expressions figées en français. Noms composés et autres locutions*. Paris: Ophrys.

Gross, Maurice and Jean Senellart. 1998. Nouvelles bases statistiques pour les mots du français. In *Journée d'Analyse Statistique des Données Textuelles (JADT), Nice 1998*, pages 335–349.

Habert, Benoît and Christian Jacquemin. 1993. Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques. *Traitement Automatique des Langues* 2:5–41.

ISO/TC 37/SC 4. 2007. Language resource managment-Lexical markup framework (LMF), ISO DIS 24613:2007. http://lirics.loria.fr/documents.html.

Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Karttunen, Lauri. 1993. Finite-State Lexicon Compiler. Tech. Rep. ISTL-NLTT2993-04-02, Xerox PARC.

Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-Level Morphology with Composition. In *Proceedings of COLING-92, Nantes*, pages 141–148.

Krstev, Cvetana, Duško Vitas, and Agata Savary. 2006a. Prerequisites for a Comprehensive Dictionary of Serbian Compounds. *LNCS* 4139:552–563.

Krstev, Cvetana, Ranka Stanković, Duško Vitas, and Ivan Obradović. 2006b. Workstation for Lexical Resources - WS4LR. In *5th International Conference on Language Resources and Evaluation, LREC'06, Genoa, Italy*, pages 1692–1697. ELRA.

Kyriacopoulou, Tita, Safia Mrabti, and Anastasia Yannacopoulou. 2002. Le dictionnaire électronique des noms composés en grec moderne. *Lingvisticæ Investigationes* 25(1):7–28.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York-London.

Oflazer, Kemal, Özlem Çetonoğlu, and Bilge Say. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Second ACL Workshop on Multiword Expressions, July 2004*, pages 64–71.

Przepiórkowski, Adam and Marcin Woliński. 2003. The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, EACL 2003*, pages 109–116.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expresions: A Pain in the Neck for NLP. *LNCS* 2276:189–206.

Savary, Agata. 2000. Recensement et description des mots composés - méthodes et applications. PhD Thesis. Université de Marne-la-Vallée.

Savary, Agata. 2005. A formalism for the computational morphology of multi-word units. *Archives of Control Sciences* 15(3):437–449.

Savary, Agata and Christian Jacquemin. 2003. Reducing Information Variation in Text. *Lecture Notes in Artificial Intelligence* 2705:145–181. Springer.

Savary, Agata, Cvetana Krstev, and Duško Vitas. 2007. Inflectional Non Compositionality and Variation of Compounds in French, Polish and Serbian, and Their Automatic Processing. *BULAG* 32:73–93.

Silberztein, Max. 1993a. *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*. Paris: Masson.

Silberztein, Max. 1993b. Les groupes nominaux productifs et les noms composés lexicalisés. *LingvisticæInvestigationes* 17(2):405–425.

Silberztein, Max. 2005. NooJ's dictionaries. In *Proceedings of LTC'05, Poznań*, pages 291–295. Wydawnictwo Poznańskie.

Villavicencio, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical Encoding of MWEs. In *Second ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, pages 80–87.

Appendix

1 Extracts of DELAC files

1.1 French DELAC, Silberztein (1993a)

cordon(N1)/bleu(A32),un:ms/-+
cousin(N32)/germain(A32),un:ms/++
toile(N21)/de/araignée,une:fs/-+
toiles(N21)/de/araignées,une:fp/--
mémoire(N21)/vive(A21),une:fs/-+

1.2 English, French and Polish DELAC, Savary (2000)

%+/+
gentleman(gentleman.N8:s) farmer(farmer.N1:s),N:s/+N
man(man.N8:s) servant(servant.N1:s),N:s/+N
%-/+
bas-relief(relief.N7:s),N:s/+N
birth date(date.N1:s),N:s/+N
man eater(eater.N1:s),N:s/+N
students' union(union.N1:s),N:s/+N
%+/-/-
date(date.N1:s) of birth,N:s/+N
%-/-
%p:p/-
%p:-/p
attorney(attorney.N1:s) general(general.N1:s),N:s/+N
battle(battle.N1:s) royal(royal.N1:s),N:s/+N
%-/+
%s:-/-
%s:p/-
%p:-/p
%p:p/p
student(student.N1:s) union(union.N1:s),N:s/+N

%+/+
cordon(cordon.N1:ms) bleu(bleu.A32:ms),N:ms/+N
cousin(cousin.N32:ms) germain(germain.A32:ms),N:ms/+N+G
mémoire(mémoire.N21:fs) vive(vif.A38:fs),N:fs/+N
%+/-/-
%p:p/-/-
%p:p/-/p
toile(toile.N21:fs) d'araignée(araignée.N21:fs),N:fs/+N

%+/+
zimne(zimny.A-ny:Mfp) nogi(noga.N4-ga:Mfp),N+AN:Mfp/+C
%+/-
%N:N/N
%C:C/C
majster(majster.N1-er:Mos)
klepka(klepka.N4-ka:Mfs),N+NN:Mos/+N+C

1.3 English, French, Polish and Serbian DELAC in Multiflex

attorney(attorney.N1:s) general(general.N1:s),NC_NXN1
bas-relief(relief.N7:s),NC_XXN
battle(battle.N1:s) royal(royal.N1:s),NC_NXN1
birth date(date.N1:s),NC_NN_NofN
gentleman(gentleman.N8:s) farmer(farmer.N1:s),NC_NXN
man eater(eater.N1:s),NC_XXN
man(man.N8:s) servant(servant.N1:s),NC_NXN
student(student.N1:s) union(union.N1:s),NC_NXN1s

cordon(cordon.N1:ms) bleu(bleu.A32:ms),NC_NXA
cousin(cousin.N32:ms) germain(germain.A32.N:ms),NC_NXAmf
mémoire(mémoire.N21:fs) vive(vif.A38:fs),NC_NXA
toile(toile.N21:fs) d'araignée(araignée.N21:fs),NC_NDN1

majster(majster.N1-er:Mos) klepka(klepka.N4-ka:Mfs),NC_NXN1
zimne(zimny.A-ny:Mfp) nogi(noga.N4-ga:Mfp),NC_AXNninv

radio-aparat(aparat.N1:ms1q),NC_2XN6

2 Extracts of DELACF files

2.1 English

attorney general,attorney general.N:s
attorneys general,attorney general.N:p
attorney generals,attorney general.N:p
bas-relief,N:s
bas-relieves,N:p
battle royal,battle royal.N:s
battles royal,battle royal.N:p
battle royals,battle royal.N:p
birth date,birth date,N:s
birth dates,birth date,N:p
date of birth,birth date,N:s # Multiflex
date of birth,date of birth,N:s # Savary (2000)
dates of birth,birth date,N:p # Multiflex
dates of birth,date of birth,N:p # Savary (2000)
gentleman farmer,gentleman farmer.N:s
gentlemen farmers,gentleman farmer.N:p
man eater,man eater.N:s
man eaters,man eater.N:p
man servant,man servant.N:s
men servants,man servant.N:p
student union,student union,N:s
students union,student union,N:s
students' union,student union,N:s # Multiflex
students' union,students' union,N:s # Savary (2000)
student unions,student union,N:p
students unions,student union,N:p
students' unions,student union,N:p # Multiflex
students' unions,students' union,N:p # Savary (2000)

2.2 French

cordon bleu,cordon bleu.N:ms
cordons bleus,cordon bleu.N:mp
cousin germain,cousin germain.N:ms
cousins germains,cousin germain.N:mp
cousine germaine,cousin germain.N:fs
cousines germaines,cousin germain.N:fp
mémoire vive,mémoire vive.N:fs
mémoires vives,mémoire vive.N:fp
toile de araignée,toile de araignée.N:fs

toiles de araignée,toile de araignée.N:fp
toiles de araignées,toiles de araignées.N:fp #Silberztein (1993a)
toiles de araignées,toiles de araignée.N:fp #Savary (2000), Multiflex

2.3 Polish

majster klepka,majster klepka.N:Mos
majstra klepki,majster klepka.N:Dos
majstrowi klepce,majster klepka.N:Cos
majstra klepkę,majster klepka.N:Bos
majstrem klepką,majster klepka.N:Ios
majstrze klepce,majster klepka.N:Los
majstrze klepko,majster klepka.N:Vos
majstrzy klepki,majster klepka.N:Mop
majstrów klepek,majster klepka.N:Dop
majstrom klepkom,majster klepka.N:Cop
majstrów klepki,majster klepka.N:Bop
majstrami klepkami,majster klepka.N:Iop
majstrach klepkach,majster klepka.N:Lop
majstrzy klepki,majster klepka.N:Vop
zimne nogi,zimne nogi.N:Mfp
zimnych nóg,zimne nogi.N:Dfp
zimnym nogom,zimne nogi.N:Cfp
zimne nogi,zimne nogi.N:Bfp
zimnymi nogami,zimne nogi.N:Ifp
zimnych nogach,zimne nogi.N:Lfp
zimne nogi,zimne nogi.N:Vfp

2.4 Serbian

radio aparat,radio-aparat.N:s1qm
radio aparata,radio-aparat.N:s2qm
radio aparatu,radio-aparat.N:s3qm
radio aparat,radio-aparat.N:s4qm
radio aparate,radio-aparat.N:s5qm
radio aparatom,radio-aparat.N:s6qm
radio aparatu,radio-aparat.N:s7qm
radio aparati,radio-aparat.N:p1qm
radio aparata,radio-aparat.N:p2qm
radio aparatima,radio-aparat.N:p3qm
radio aparate,radio-aparat.N:p4qm
radio aparati,radio-aparat.N:p5qm
radio aparatima,radio-aparat.N:p6qm
radio aparatima,radio-aparat.N:p7qm

radio aparata,radio-aparat.N:w2qm
radio aparata,radio-aparat.N:w4qm
radio-aparat,radio-aparat.N:s1qm
radio-aparata,radio-aparat.N:s2qm
radio-aparatu,radio-aparat.N:s3qm
radio-aparat,radio-aparat.N:s4qm
radio-aparate,radio-aparat.N:s5qm
radio-aparatom,radio-aparat.N:s6qm
radio-aparatu,radio-aparat.N:s7qm
radio-aparati,radio-aparat.N:p1qm
radio-aparata,radio-aparat.N:p2qm
radio-aparatima,radio-aparat.N:p3qm
radio-aparate,radio-aparat.N:p4qm
radio-aparati,radio-aparat.N:p5qm
radio-aparatima,radio-aparat.N:p6qm
radio-aparatima,radio-aparat.N:p7qm
radio-aparata,radio-aparat.N:w2qm
radio-aparata,radio-aparat.N:w4qm
radioaparat,radio-aparat.N:s1qm
radioaparata,radio-aparat.N:s2qm
radioaparatu,radio-aparat.N:s3qm
radioaparat,radio-aparat.N:s4qm
radioaparate,radio-aparat.N:s5qm
radioaparatom,radio-aparat.N:s6qm
radioaparatu,radio-aparat.N:s7qm
radioaparati,radio-aparat.N:p1qm
radioaparata,radio-aparat.N:p2qm
radioaparatima,radio-aparat.N:p3qm
radioaparate,radio-aparat.N:p4qm
radioaparati,radio-aparat.N:p5qm
radioaparatima,radio-aparat.N:p6qm
radioaparatima,radio-aparat.N:p7qm
radioaparata,radio-aparat.N:w2qm
radioaparata,radio-aparat.N:w4qm

3 English, French, Polish and Serbian inflectional graphs in Multiflex

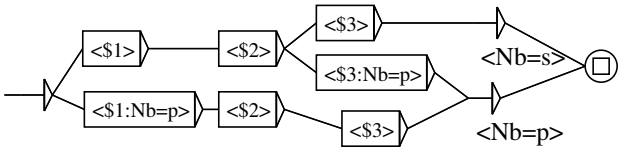


FIGURE 5 *Multiflex* inflection graph NC_NXN1 for *attorney general* and *battle royal* in English

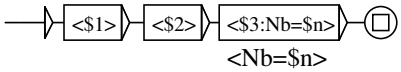


FIGURE 6 *Multiflex* inflection graph NC_XXN for *bas-relief* and *man eater* in English

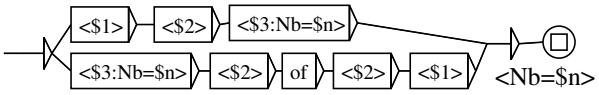


FIGURE 7 *Multiflex* inflection graph NC_NN_NofN for *birth date* in English

FIGURE 8 *Multiflex* inflection graph NC_NXN for *gentleman farmer* and *man servant* in English

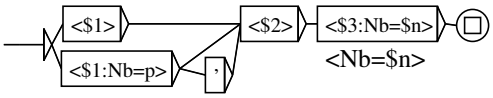


FIGURE 9 *Multiflex* inflection graph NC_NXN1s for *student union* in English

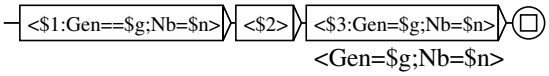


FIGURE 10 *Multiflex* inflection graph NC_NXA for *cordon bleu* and *mémoire vive* in French

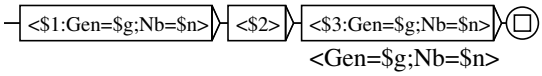


FIGURE 11 *Multiflex* inflection graph NC_NXAmf for *cousin germain* in French

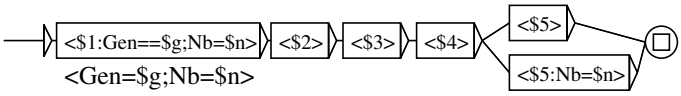


FIGURE 12 *Multiflex* inflection graph NC_NDN1 for *toile d'araignée* in French

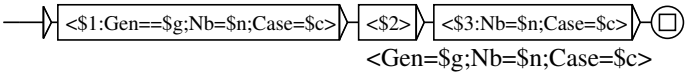


FIGURE 13 *Multiflex* inflection graph NC_NXN1 for *majster klepka* in Polish

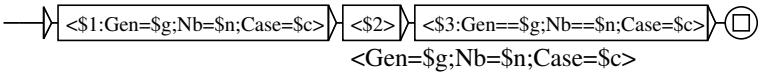


FIGURE 14 *Multiflex* inflection graph NC_AXNninv for *zimne nogi* in Polish

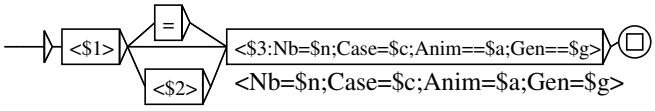


FIGURE 15 *Multiflex* inflection graph NC_2XN6 for *radio-aparat* in Serbian