A Formalism for the Computational Morphology of Multi-Word Units

Agata Savary

Université François Rabelais,
Laboratoire d'Informatique,
IUT de Blois, 3 pl. Jean-Jaurès, 41029 Blois cedex, France
agata.savary@univ-tours.fr

Abstract

Multi-word units (MWUs) are linguistic objects placed on the frontier between morphology and syntax. A reliable computational treatment of their inflectional morphology requires a fine-grained grammar-based approach allowing a description of general large-coverage phenomena as well as of lexicalized counter-rules. We propose a formalism that answers these requirements. Due to a graph-based description and a simple unification algorithm it allows to compactly and exhaustively describe the inflection paradigm of a MWU in terms of its component words' morphology as well as of some regular-language patterns.

1. Introduction

Multi-word units (MWUs), are linguistic objects placed on the frontier between morphology and syntax: their compound structure suggests productivity (repetitiveness of a phenomenon for units sharing the same structure) which can hardly be processed without a grammar-based approach, while some of their morphological, syntactic and semantic properties exclude their processing merely in terms of the properties of their constituents.

The inflectional morphology of MWUs is one of the areas in which a grammar-based approach must necessarily be accompanied by a set of lexicalized counter-rules in order for the automatic morphological processing to be reliable. 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 *battle cry, battle royal* and *battle of nerves* in English not only do we need to generate the plural of *battle, royal* and *cry* but also to know how different inflected forms of these constituents combine (*battle cr<u>ies</u>*, *battle royal<u>s</u>* or *battle<u>s</u> royal*, *battle<u>s</u> of nerves*, but not **battle<u>s</u> cr<u>ies</u>*, **battle<u>s</u> royal<u>s</u>*, **battle<u>s</u> of nerve<u>_</u>*).

Most of the former approaches to inflectional morphology of MWU are based on a single word-to-word basis (e.g. lemmatizing or stemming of constituent words, generation of all combinations of all inflected forms of the constituents) or on ad hoc rules. Such approaches prove to be robust but they suffer either from excessive generalizations or from overlooking of exceptions.

We propose a formalism that allows to exhaustively and compactly describe the inflection paradigm of a MWU in terms both of its component words' morphology and of some regular-language patterns. This description may be used in the computational analysis and generation of application-dependent MWUs such as compounds, frozen expressions, complex terms, multi-word named entities, etc.

2. Multi-word units. Scope of the paper

Multi-word units (MWUs) encompass a bunch of hard-to-define and controversial linguistic objects (cf. (Habert and Jacquemin, 1993), (Corbin, 1992)). Their numerous linguistic and pragmatic definitions (e.g. (Benveniste, 1974), (Downing, 1977), (Levi, 1978), (Gross, 1990), (Silberztein, 1993b), (Gross, 1996), (Cadiot, 1992), (Derwojedowa and Rudolf, 2003)) invoke three major points:

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

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

We define the scope of our work pragmatically: we consider a MWU as a contiguous sequence of *graphical units* which, for some application-dependent reasons, has to be listed, described (morphologically, syntactically, semantically, etc.) and processed as a unit.

Further, we limit ourselves to the formal description of the inflectional paradigms of MWUs which requires to answer the following questions:

- What is the MWU's morphological class (noun, adjective, etc.)?
- What are the possible combinations of inflection features for the MWUs? E.g. in Polish *wybory powszechne* (*general election*) does not have a singular form (although its head word *wybory* does).
- What are the inflectional characteristics (base form, morphological class, inflection paradigm) of the single constituents of the MWU? E.g. in French, *porte* (*door*) is an uninflected verb in *porte-avion* (*aircraft carrier*) while it is an inflected noun in *porte-fenêtre* (*French window*) which takes an *s* in plural (*porte<u>s</u>-fenêtre<u>s</u>*).
- How should we combine the inflected forms of the single constituents in order to generate the inflected forms of the whole compound? E.g. to inflect *battle of nerves* and *battle cry* in number we need to inflect the first and the last constituent, respectively.

3. State of the art

Most of the existing approaches to the inflectional morphology of MWUs have been developed in view of their direct application to either morphological analysis or morphological generation.

One of the goals of morphological analysis is to determine the base forms and/or morphological features of inflected forms of MWUs (e.g. *customs officers* → [*customs officer*, plural]). In some domains of information retrieval, as term acquisition and automatic indexing, the conflation of morphological, syntactic and semantic variants of extracted compound terms proved to be crucial (Savary and Jacquemin, 2003). In these domains, a frequent approach to inflectional analysis of MWUs is stemming or lemmatizing of all their component words, e.g. (Daille, 1994), and (Dillon and Gray, 1983). This method may give erroneous results in some cases.

First, the lemma of a compound may contain words that are not lemmas themselves. For instance, the singular of *customs barriers* is *customs barrier* instead of **custom barrier*. One may agree that an incorrect lemma such as **custom barrier* is an abstract reference form, as in the case of stemming of single words (e.g. *deni* for *denied*). However, this is inconvenient for two reasons: it makes any human post-processing tedious, and it may result in spurious conflations with free form sequences, as in <u>*custom barrier*</u> *bags for protection of equipment*.

Second, a single word may carry an inflection mark of a compound although this word has no inflection as an individual lexical item. For instance, in *court <u>martials</u>, good-for-<u>nothings</u>, stand-<u>bys</u>,* etc. the underlined words cannot be lemmatized by a standard dictionary-based morphological analyzer. Such cases may be handled by a stemmer though.

The lemmatizing of compounds in French is even more complex due to the gender inflection of adjectives, as well as to the important number of compounds with a non-standard nominal construction, e.g. *un porte-avions* (*aircraft carrier*), *une deux-chevaux* (a car with a two horse-power engine).

A reliable inflection generation has two characteristics:

- the correctness: each generated form is correct
- the exhaustivity: all existing inflection variants may be generated (e.g *attorney general* in plural yields *attorney generals, attorneys general*, and *attorneys generals*)

The first issue is crucial for instance in the computer aided translation. If we automatically replace a source language term by its target language equivalent, a good strategy is to possibly preserve the inflection features of the original term (e.g. *French windows* → *portes-fenêtres*). This target form has to be fully correct in order to avoid human post-processing.

The second issue is important in two kinds of applications: dictionary-based morphological analyzers, e.g. (Paumier, 2003), (Silberztein, 1993a), and (Evans et al., 1991), and variant generators for phrase indexing (Sparck-Jones and Tait, 1984), in which all inflection variants of MWUs are generated beforehand, and then identified through a straightforward string matching in the corpus.

Previously, we proposed (Savary, 2000) a formalism which allows to avoid the above problems (but requires a good deal of manual lexicographic treatment). It is based on the detection of the so-called *characteristic constituents* (CC) of a MWU, i.e. of its headword and all other words that morphologically agree with the headword, e.g. the first two constituents in the French term <u>*réseau numérique*</u> *par satellite* (*digital satellite network*). The inflectional paradigm of a MWU is determined implicitly in the case of a regular inflection: the MWU has the same inflection features as its CCs, and in order to inflect it we need to inflect all these constituents and only them (*réseau<u>x</u> numérique<u>s</u> par satellite*). Thus, the morphological description only needs to point at the CCs. In the case of irregular inflection some special rules explicitly describe the exception forms as combinations of particular forms of the (characteristic or non characteristic) constituents, e.g. *franc-maçon* → *franc<u>s</u>-maçon<u>nes</u>* (the feminine plural of the whole MWU is obtained from the masculine plural of *franc* and the feminine plural of *maçon*).

That formalism has some disadvantages. First, the base form of a MWU appears as its particular inflected form to which some inflection values are attributed, e.g. the base form *battle royal* is described as singular, *wybory powszechne* (*general election*) as plural, etc. However, linguistically speaking, it is preferable to adopt the notion of lemma as an *abstract* representative of the set of inflected forms, i.e. possessing no inflection features itself. This is for example the case in the morphological description of simple words by (Courtois, 1990) where each lemma is only attributed an inflection code, i.e. a pointer to its paradigm's description (e.g. *numérique,A31*).

Second, the inflection paradigm of a MWU is not homogeneous. Some parts of it appear on the entry level while the rest is factorized and placed in the preamble. It would be preferable to detach the paradigm description from the particular data representation as it is the case for simple words in the DELA convention (Courtois, 1990).

Third, the identifiers for inflection features and values may only be single unique characters (e.g. *N* for number, *s* for singular, etc.). This constraint may be too strong, especially for highly inflected languages.

Two other approaches to explicit and exhaustive description of the inflectional morphology of MWUs have been undertaken in the context of automatic generation of DELA-like electronic dictionaries for compounds. 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 combination of inflected forms of the particular constituents. This model suffers from several drawbacks. Firstly, the definition of a graphical unit is fixed as it results from the underlying model for simple words. Secondly, there is a lack of flexibility as to the possible combinations of inflection features admitted by the filters (e.g. it is impossible to express that a particular form of one constituent combines only with a particular form of another constituent). Thirdly, the rules for determination of inflection features of a MWU are heterogeneous and non generic, which makes the formalism hardly adaptable to a different language. Finally, the presence or the absence of separators (which may be relevant to inflection) cannot be expressed by restriction filters.

(Silberztein, 2005) suggests a uniform formalism for the inflection of both compounds and simple words. It is based on actions to be performed on the suffixes of particular graphical constituents of an entry. The main problem in this model is that a simple word's morphology has to be described as many times as it appears in an MWU at a position subject to inflection. E.g., 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 reproach becomes essential in case of highly inflected languages like e.g. Slavic languages. On the other hand, the advantage of this formalism, as opposed to ours presented below, is the possibility of a unique paradigm description for compounds of different lengths, e.g. *bag of tricks, balance of payment, deficit member of the opposite sex*, etc.

4. New formalism

In our new approach we try to avoid the disadvantages of the former one (Savary, 2000) by a better modularity and explicitness. We first propose a general morphological description on the language level, which should list all inflectional categories (gender, case,...) and values (masculine, singular, nominative,...). For instance, for Polish such description might be as follows:

Nb:sing,pl
Gen:pers_masc,anim_masc,inanim_masc,fem,neut
Case:Nom,Gen,Dat,Acc,Inst,Loc,Voc
...

Thus, we allow the category and value identifiers to be strings rather than single characters.

4.1. Underlying morphological module for simple words

Clearly, in order to be able to inflect MWUs we first have to know how their constituents inflect. Numerous approaches to the morphology of simple words exist. Rather than propose our own treatment of simple words we suppose that the description of their morphology is done by an existing external module. Moreover, we wish our formalism to work with possibly many such modules. Thus, we have defined a set of interface constraints that need to be satisfied in order for an external module to be applicable to our formalism.

Firstly, it has to share the alphabet as well as the morphological classes (noun, adjective,...), categories (gender, case,...) and values (masculine, singular, nominative,...) with our system.

Secondly, it has to allow, for each simple word and a set of relevant inflection values, the generation of the corresponding inflected forms (one or more, in case of variants), e.g. for *ręka* (*hand*):

(*ręka*,[Nb=pl,Gen=fem,Case=Inst]) → {*rękami*, *rękoma*})

Thirdly, it has to provide a clear definition of a graphical unit that it considers indivisible. The notion of a graphical unit varies from one approach to another (e.g. *c'est-à-dire* is considered as a sequence of 7 tokens with some systems, while it is an indivisible token with others). Our formalism may work with any of such definitions under the condition that the underlying module for simple words handles the inflection (if any) of the units that it has itself defined. Note however that the inflection paradigms of MWUs depend on that definition (as show below) so their design is different with different underlying single words' modules. In the following examples we will admit that a graphical unit is either a contiguous sequence of alphabet characters (e.g. *dire*) or a single non alphabet character (e.g. a hyphen, a dot, a digit, etc.). If that definition changes the following description of the formalism remains the same up to examples.

The above definition of interface between the morphological description of simple words and the one of MWUs allows a better modularity and independence of one another. The latter doesn't need to know how inflected forms of simple words are described, analyzed and generated. It only requires a set of correct inflected forms of a MWU's constituents. Conversely, the former formalism knows nothing about how the latter one combines the provided forms to produce multi-word sequences.

4.2. Inflection graphs for MWUs

The inflectional paradigm of a MWU is represented as a graph in which each path corresponds to one or more inflected forms. For example, figure 1 represents the inflection paradigm of MWUs that inflect like *battle royal*. Each of the numerical variables $1, $2, and $3 represents the corresponding constituent (here: *battle*, space, and *royal*). If the variable appears alone in a box, the constituent is to be the same as in the base form of the MWU. For instance, <$3> in the uppermost path means that the unit *royal* is to be recopied as such. If the variable is accompanied by a set of category-value equations, the constituent has to be inflected to the required form. E.g. <$3:Nb=p> means that the plural form of *royal* is needed.

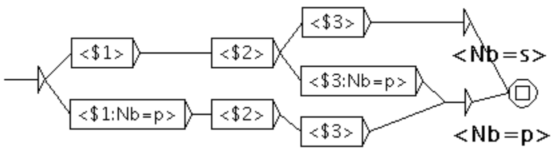

Figure 1: Inflection graph for *battle royal*

The generation of inflected forms starts with the leftmost right arrow and ends with the final encircled box. Each time we come to a node we perform the (possibly empty) action contained in the box (a recopy of a constituent, or a generation of its particular inflected form), and we accumulate the morphological features contained under the box. The total of the accumulated features should result in the complete morphological description of the inflected form. For instance, the exhaustive exploration of the graph on figure 1 with *battle royal* results in the set of 3 forms:

battle royal <Nb=s>
battle royals <Nb=p>
battles royal <Nb=p>

Note that this description is independent of the way we generate inflected forms of single words because we suppose that this problem is handled by an existing external morphological system for single words.

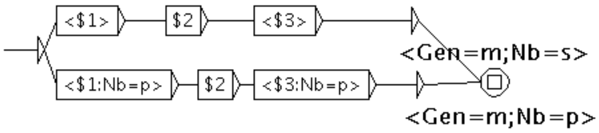

Figure 2: Inflection graph for *bateau-mouche*

In an inflection paradigm of a MWU, each constituent is accompanied only by those morphological categories which it should inflect for. The categories that remain unchanged don't have to be mentioned. For instance, in *bateau-mouche* in French (a Paris-style river-boat), both noun constituents have their gender set but they inflect in number: *bateaux-mouches*. That's why on figure 2 containing the inflection graph for this MWU, the corresponding boxes contain value assignments for number only. Note that both constituents may or may not agree in gender, here *bateau* is masculine while *mouche* is feminine.

4.3. Unification Variables

An important feature of our formalism are **unification variables**. They are introduced by the dollar sign followed by an identifier which may contain any number of characters, e.g. *$g1*, *$num_10*, *$c*, etc. For example, figure 3 shows a graph roughly equivalent to the one on figure 2 in the sense that it allows to generate the same inflected forms for the same MWUs. However, this time a single path represents both the singular and the plural form. That is possible due to the unification variable *$n* which may be instantiated to any value of the domain of its category (*Nb*), here *$n=s* or *$n=p*. The instantiation is unique for all elements on a path: if we fix the singular value for the first constituent the same value has to be set for the third one, as well as for the whole MWU. Similarly, if we fix *$n* to *p* while processing the first node it has to remain *p* until the end of the path.

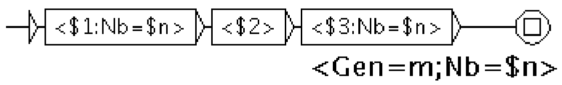

Figure 3: Inflection graph for *bateau-mouche* with a unification variable

The inflection graph on figure 3 applies to most kinds of French compounds of types *Noun Noun* and *Noun Adjective* (*bateau-mouche, ange gardien, circuit séquentiel*, etc.) which are of **masculine** gender. That is because the output of the final node contains *Gen=m*. For all compounds of the same types but of feminine gender, e.g. *main courante, moissoneuse-batteuse*, etc., a new graph has to be created which is identical to figure 3 up to the final output containing *<Gen=f;Nb=$n>*. That is not very intuitive since *circuit séquentiel* and *main courante* inflect in the same way, in the sense that in both cases we need to put the first and the last constituent to plural in order to obtain the plural form of the whole MWU.

That's why another type of instantiation for unification variables has been introduced. It is accompanied by a **double equal sign** (==) (as opposed to the **single equal sign** = as for *$n* on figure 3). If a unification variable is assigned to a category by this symbol then it **inherits** the value of this category from the corresponding constituent, as it appears in the lemma of the MWU. For instance, figure 4 contains a graph describing the inflected forms for both masculine and feminine French compounds of types *Noun Noun* and *Noun Adjective*. Its first box contains the double assignment of the gender to variable *$g* which means that this variable has its value fixed to the gender value of the first constituent. For *bateau-mouche* it is fixed to masculine because *bateau* is masculine, while for *main courante* it is fixed to feminine.

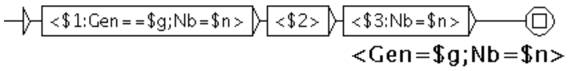

Figure 4: Inflection graph for *bateau-mouche* with two types of instantiation

Note that the double assignment, contrary to the single assignment, no longer means that the variable is to be instantiated to all values of the corresponding category domain. It has a unique value all through the path on which it appears, even if it is concerned by another, single, assignment somewhere else on the same path. For example, on figure 4 the final output contains *Gen=$g* but *$g* may only take one value determined by the first constituent.

Unification variables are particularly useful in highly inflected languages. For example, in Polish most nouns inflect for number (2 values) and case (7 values), which implies at least 14 different forms (if variants and syncretic forms are distinguished). This score is even higher for adjectives which inflect for number, case and gender (3 till 9 values, according to different approaches). If no unification mechanism were available each of these numerous forms would have to be described by a separate path in the graph. The use of unification variables allows to dramatically reduce the size of the graph (to one path only in most cases).

For example, figure 5 shows the graph for Polish compounds that inflect like *pranie mózgu* (*brainwashing*) or *powożenie koniem* (*horse coaching*). Their third constituent has its case fixed (most often to genitive or instrumental). Their first and third constituent inflect in number independently from each other (*pranie mózgów*, *prania mózgu*, *prania mózgów*, etc.). That's why either of them has a different unification variable for number inflection (*$n1* and *$n2*). The three variables $n1, $n2, and $c may be instantiated to any value from their respective domains (*{sing,pl}*, *{sing,pl}*, and

{Nom,Gen,Dat,Acc,Inst,Loc,Voc}). The whole MWU inherits its gender, number and case from its first constituent. Its gender is fixed (*Gen==$g*) while its number and case are instantiated to any of the 14 possible combinations. The single path in this graph would have to be replaced by 28 different ones if the use of unification variables were not allowed.

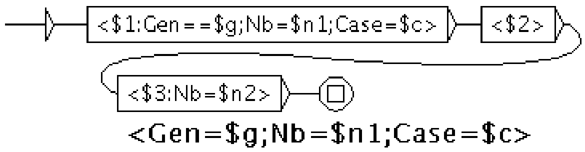

Figure 5: Inflection graph for *pranie mózgu*

4.4. Orthographic and Other Variants

Our formalism allows for any constituent to be omitted or moved within different inflected forms if there is a need for that. It also enables the insertion of extra graphical units which do not appear in the base form of the MWU. This allows to extend an inflection paradigm to a more general variation description, e.g. orthographic or, partly, syntactic variation (see (Jacquemin, 2001) for an extensive study on term variation). For example, in English, *student union* appears in corpus also as *students union*, and *students' union*, in singular or plural in each case. Our formalism allows to include both types of variation in one description (cf. figure 6).

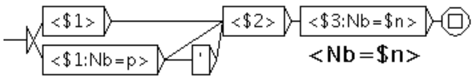

Figure 6: Inflection graph for *student union*

Figure 7 shows an example in which, additionally to the insertion of a new constituent, the order of constituents may be reverted. The upper path allows to generate e.g. *birth date* and *birth dates* while the lower one represents the syntactic variants of the previous forms: *date of birth* and *dates of birth*.

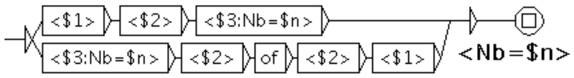

Figure 7: Inflection graph for *birth date*

5. Complete interface example

Let's consider an interface of our formalism with the DELA convention (Courtois, 1990) implemented in the Unitex system (Paumier, 2003). This system describes the inflectional morphology of a simple word via an inflection code attributed to its lemma. For instance, in English, nouns taking an *s* or an *ies* in plural obtain codes *N1* and *N3*, respectively, as shown on figures 8 and 9. Each graph node contains the action to be performed on the lemma suffix (here, the concatenation of an *s*, or the deletion of the last letter and the concatenation of an *ies*), while the corresponding inflection features are placed under the node.

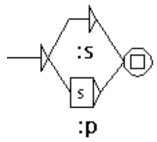

Figure 8: Inflection graph *N1* for English simple words

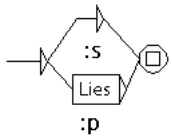

Figure 9: Inflection graph *N3* for English simple words

Let's admit that the following MWUs are to be inflected: *battle royal, student union*, and *birth date*. Their constituents subject to inflection must first be morphologically identified, i.e. tagged with information that is sufficient for Unitex to generate any of their inflected forms: their lemmas, inflection codes and morphological features. Then, inflection codes have to be attributed to MWUs. The whole results in the following description where *NC_NN_NofN* and *NC_NsNs* are the graphs on figures 7 and 1, respectively:

birth date(date.N1:s),NC_NN_NofN
notary(notary.N3:s) public(public.N1:s),NC_NsNs

If we run the graphs on this description and format the generated forms according to the Unitex standards we obtain the following list of inflected forms, accompanied by theirs lemmas, codes, and inflectional values:

date of birth,birth date.NC_NN_NofN:s
dates of birth,birth date.NC_NN_NofN:p
birth date,birth date.NC_NN_NofN:s
birth dates,birth date.NC_NN_NofN:p
notaries public,notary public.NC_NsNs:p
notary public,notary public.NC_NsNs:s
notary publics,notary public.NC_NsNs:p

6. Critical remarks and perspectives

Some refinements of the proposed formalism are worth examination. Firstly, the formalism is based on the assumption that a clear-cut model of the language's morphological values is possible as shown for Polish at the beginning of section 4. However, it is not obvious that, for some languages that we have not yet studied, such a disjunctive division into categories and values is always possible and natural. Some productive cross-references between categories may exist that are impossible to express on the level of the general language description in our formalism. We hope however that due to the regular language patterns describable by infection graphs, most of such cross-references may be expressed at least on the level of inflectional paradigms.

Secondly, it would be interesting to be able to *recursively* describe the inflection of MWUs in function of their constituents. Let's assume, in French for instance, that the inflection of *liaison numérique* (*digital link*) has already been described in the MWU morphological module. Then we may consider *liaison numérique par satellite* (*satellite digital link*) as a compound composed of five units: *liaison numérique*, a space, *par*, a space, and *satellite*. The generation of the inflected forms of this compound could refer to the inflection of *liaison numérique* as a whole instead of *liaison* and *numérique* separately.

Thirdly, our formalism has been designed for the purpose of the generation of *all* inflected forms of a MWU. That is because, it has been originally inspired by the DELA formalism (Courtois, 1990) implemented by the Unitex system (Paumier, 2003) in which inflection codes for simple words are mainly used in order to generate a DELAF (electronic dictionary of inflected forms of simple words) out of a DELAS (electronic dictionary of simple words' lemmas). However, it would be interesting to be able to use the same formalism for other purposes, e.g. for generation of only one particular inflected form. Inflection graphs, as presented above, are not optimal for this kind of usage.

Finally, it is necessary to note that in our approach the description is fully lexicalized (thus, potentially labor intensive) in the sense that an inflection paradigm has to be assigned to each MWU in order to fully describe its inflected forms. An important point that hasn't been addressed up to now is *how* inflection graphs are assigned to MWUs. We think that this process may be automatic to a big extent. It could rely on a *grammar*, i.e. some language-dependent inflection regularities such as: 'in Polish, if a MWU is composed of a noun in nominative followed by a noun in genitive then in most cases the second noun is never inflected', as in *świadectwo dojrzałości* (*secondary school certificate*). However, a human post-processing will always be necessary in order to detect exceptions to such general rules such as *pranie mózgu* (cf section 4.3.). The expressive power of the presented graph formalism allows to exhaustively describe these exceptions. The development of such a grammar-based controlled assignment of graphs to MWUs is our major perspective.

7. Conclusion

We have presented a proposal of a formalism for the description of inflectional paradigms of multi-word units. It is based on acyclic graphs fully and precisely describing all inflected forms of MWUs, independently of the morphological system of simple words to be used. It provides a compact representation of numerous inflected forms due to a unification mechanism. It also allows for a joint description of different types of variation.

8. Acknowledgments

The author is grateful to Marcin Woliński, Adam Przepiórkowski, Anna Kupść and Duško Vitas for their critical remarks on previous versions of the paper.

9. References

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

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

Corbin, D., 1992. Hypothèthes sur les frontières de la composition nominale. *Cahiers de grammaire*, 17.

Courtois, B., 1990. Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87.

Daille, B., 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Ph.D. thesis, Université Paris 7.

Derwojedowa, M. and M. Rudolf, 2003. Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu. *Poradnik językowy*, 5:39–49.

Dillon, M. and A.S. Gray, 1983. FASIT: A fully automatic syntactically based indexing system. *Journal of the American Society for Information Science*, 34(2):99–108.

Downing, P., 1977. On the Creation and Use of English Compound Nouns. *Language*, 153(4).

Evans, D.A., G.-W. Kimberly, M. Hart, R.G. Lefferts, and I.A. Monarch, 1991. Automatic indexing using selective NLP and first-order thesauri. In *The Proceedings, Intelligent Multimedia Information Retrieval Systems and Managment (RIAO'91), Barcelona*.

Gross, G., 1990. Définition des noms composés dans un lexique-grammaire. *Langue Française*, 87.

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

Habert, B. and Ch. Jacquemin, 1993. Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *Traitement Automatique des Langues*, 2.

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

Kyriacopoulou, T., S. Mrabti, and A. Yannacopoulou, 2002. Le dictionnaire électronique des noms composés en grec moderne. *LingvisticæInvestigationes*, 25(1):7–28.

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

Paumier, S., 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée.

Savary, A., 2000. *Recensement et description de mots composés – méthodes et applications*. Ph.D. thesis, Université de Marne-la-Vallée.

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

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

Silberztein, M., 1993b. Les groupes nominaux productifs et les noms composés lexicalisés. *LingvisticæInvestigationes*, 17(2).

Silberztein, M., 2005. NooJ's dictionaries. In *The Proceedings of the 2nd Language & Technology Conference (LTC'05), Poznań, April 21-23, 2005*.

Sparck-Jones, K. and J.I. Tait, 1984. Automatic search term variant generation. *Journal of Documentation*, 40(1):50–66.