Towards 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 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 cries, battle royals* or *battles royal, battles of nerves*, but not **battles cries*, **battles royals*, **battles of nerve*).

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 generalisations 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 and goal 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 ((Benveniste, 1974), (Downing, 1977), (Levi, 1978), (Bauer, 1983), (Gross, 1990), (Anscombre, 1990), (Silberztein, 1993), (Gross, 1996), (Cadiot, 1992)) invoke three major points:

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

However, the basic notions (a word, a reference, 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, semanticaly, etc.) and processed as a unit. Given a finite alphabet for the given language, a *graphical unit* is either a contiguous sequence of alphabet characters or a single non-alphabet character.

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 (nous, adjective, etc.) and thus what inflection categories (number, gender, case, etc.) are relevant to it? (Przepiórkowski and Woliński, 2003) argue for a morphosyntactically motivated definition of morphological classes: a morphological class should fully determine the inflection categories the word inflects for as well as those that are lexically fixed for the word, e.g. in French, a noun has a gender and it inflects for number only; in Polish, a noun has a gender and inflects for number and case.
- What are the exceptions to the inflection categories determined above? E.g. in Polish *wybory powszechne* (*general election*) is a compound noun but it doesn't 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 (*portes-fenêtres*).

- 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 consituent, 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), (Dillon and Gray, 1983), and (Arampatzis et al., 1998). 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 *custom barrier 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 martials*, *good-for-nothings*, *stand-bys*, 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 horsepower 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 *notary public* in plural yields *notary publics*, *notaries public*, and *notaries publics*)

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, 2003), 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 *réseau numérique 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 (*réseaux numériques par satellite*). Thus, the morphological description only needs to point at the CCs. The MWU are described by groups, as below:

```
#+/+/-/-
réseau(réseau.N3:ms)  numérique(numérique.A31:ms)
par satellite,N+NAPrepN:ms/+N
réseau(réseau.N3:ms)  mondial(mondial.A32:ms)  de
télécommunication,N+NAPrepN:ms/+N
liaison(liaison.N21:fs)  multiple(multiple.A31:fs)  par
satellite,N+NAPrepN:fs/+N
```

Each group contains the MWUs having the same number of constituents and the same positions of their CCs, which are described in the preamble of the group (+/+/-/-). For each CC, its base form (*réseau*), the code of its inflection paradigm (*N3*), and its inflection features (*ms*) are listed. Additionally, the morphological class and type (*N+NAPrepN*) of the whole MWU are given together with its inflection features (*ms*) and the inflection categories it inflects for (*+N* for number inflection).

In the case of irregular inflection some special lines are added to the preamble as is the case below where no CCs are present and two possible plural forms are created by inflecting either the first or the second constituent while leaving the other one intact.

```
#-/-
#p:p/-
#p:-/p
battle(battle.N1:s) royal(royal.N1:s),N+NN:s/+N
```

The above formalism also includes a simple inheritance mechanism: the inflection values that are to be attributed to an inflected form of a MWU are the same as those of its characteristic constituents. It also applies a simple unification mechanism. Consider the following description:

#+/-
#C:C/C
człowiek(człowiek.N51:Mos)
kot(kot.N73:Mzs),N+NN:Mos/+C

It says that only the first constituent is characteristic, however the case inflection concerns both contintuents. In the second line of the preamble the case indentifier *C* functions as a unification variable that may be instantiated by any value of the domain described in a separate file (*C:M,D,C,B,I,L,W* means that there are 7 possible case values in Polish). All appearences of the variable get instantiated identically, thus, e.g. the locative (*L*) of the whole entry, *człowieku kocie*, (*cat man*) is obtained by concatenation of the locative forms of both constituents.

The above 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 (e.g. *+N*) while the rest is factorised 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 (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.

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 suppose that, for a given natural language, an external morphological system for simple words exists which shares the alphabet as well as the morphological classes (noun, adjective,...), categories (gender, case,...) and values (masculine, singular, nominative,...) with our system. Their description for Polish might be done 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 class, category and value identifiers to be strings rather than single characters.

The simple words' morphological system mentioned above 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*})

A MWU is graphically defined as a concatenation of graphical units, i.e. either contiguous sequences of alphabet characters (e.g. *dom*, *aujourd*, *NATO*) or single non-alphabet characters (e.g. a space, a hyphen, a dot, ...). Each non-alphabet character is considered as an independent MWU's constituent, thus e.g. *Athens '04* contains five constituents: *Athens*, space, apostrophe, digit 0, and digit 4.

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 *battle royal*. Each of the numerical variables $1, $2, and $3 represents the corresponding constituent (*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. If the variable is accompanied by a set of feature-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. Each path in the graph starts at the leftmost right arrow and ends at the final encircled box.

Figure 1: Inflection graph for *battle royal*

Note that for each constituent only those feature-value equations need to be provided that correspond to morphological categories the word's class inflects for, and not those the class has set. For example, figure 2 contains the graph for French MWUs that inflect like *bateau-mouche* (a Paris-style river boat, *bateaux-mouches* in plural). The first and the third constituent are nouns so they have their gender set. Thus, the corresponding boxes contain the value assignment for number only. Note that both constituents may or may not agree in gender, here *bateau* is masculine while *mouche* is feminine. The rightmost box of each path is the one that determines the full set of inflection values for the MWU: both the fixed and the variable ones. They may be distinguished by two types of assignments: double assignments (==) for fixed values (e.g. *Gen==masc*), and single assignments (=) for the categories the word inflects for (e.g. *Nb=s*).

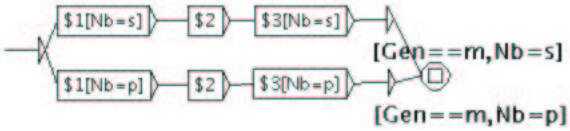

Figure 2: Inflection graph for *bateau mouche*

Inflection graphs may contain unification variables. For example, figure 3 shows a graph equivalent to the one on figure 2. Both the singular and the plural forms are represented here by the two possible instantiations of the unification variable *$n*, i.e. *$n=s* and *$n=p*. The instanti-

ation is unique for all elements of one 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.

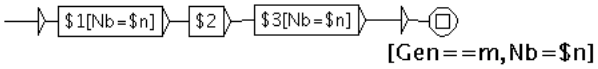

Figure 3: Inflection graph for *bateau mouche* using unification variables

Unification variables are particularly usefull 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 4 shows the graph for Polish words that inflect like *pranie mózgu* (*brainwashing*). Their third constituent is always in genitive. Their first and third constituent inflect in number independently from each other (*pranie mózgów*, *prania mózgu*, *prania mózgów*). That's why either of them has a different unification variable for number inflection (*$n1* and *$n2*). The three inflection 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}*, respectively). The whole MWU has its gender fixed to neuter and inherits its case and number from the first constituent. The single path in this graph would have to be replaced by 28 different ones if the use of unification variables were not allowed.

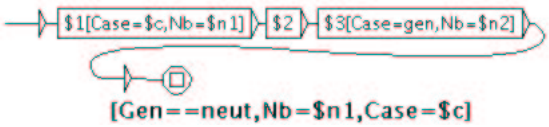

Figure 4: Inflection graph for *pranie mózgu*

Note that our formalism allows for any constituent to be omitted or moved within different inflected forms if there is a need for that. It also allows for 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 5).

5. Interface with the morphological system of simple words

In order for our formalism to be applicable to morphological analysis and generation of MWUs, an adequate

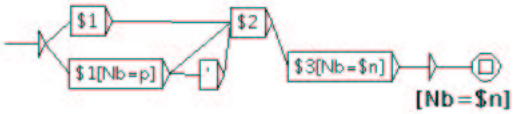

Figure 5: Inflection graph for *student union*

system for morphology and contextual disambiguation of simple words must exist which provides several automatical or semi-automatical functionalities:

- For a given simple inflected form F, and a given context (i.e. the MWU which contains F) identify all possible lemmas, the corresponding inflection paradigms, and inflection features relevant to this context. For example, in French for the form *vive* and the context *mémoire vive* (*read-only memory*) the lemma is *vif*, the inflection features are *[Gen=f,Num=s]*, and the inflection paradigm might be *N23*.[1] These elements don't have to be communicated to the MWU module but are necessary for the correct implementation of the functionality below.

- For a given lemma, inflection paradigm and a set of inflectional values, generate all corresponding inflected forms. For example, in Polish, for the lemma *ręka* (*hand*), for the inflectional paradigm that could be *N488*, and for inflection values *[Case=inst]*, the corresponding set of inflected forms would be {*ręką*, *rękami*, *rękoma*}.

This definition of interface between the morphological system for simple words and the one for 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, analysed and generated. It only requires a set of correct inflected forms of a MWU's constituents. Conversely, the former system knows nothing about how the latter one combines the provided forms to produce multi-word sequences.

6. Grammar-based vs. lexicalized decription

In our approach, an inflection paradigm is assigned to each MWU in order to fully describe its inflected forms. In this sense, the description is fully lexicalized. However, 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

[1] Note that the contextual disambiguation in this case is not equivalent to a classical tagging module, as the context of a simple word is limited to the MWU it appears in, without any larger corpus occurence of this MWU.

general rules such as *pranie mózgu* (cf section 4.). The expressive power of the presented graph formalism allows to exhaustively describe these exceptions. The development of such a grammar-based controlled description for the assignment of graphs to MWU is our major perpective.

7. Critical remarks and perspectives

Some refinements of the proposed formalism are worth examination. First, the definition, inspired by (Paumier, 2003), of an indivisible graphical unit we mention in section 4. is arbitrary. If we wish to easily develop an interface with any morphological system for simple words meeting the requirements mentioned in section 5., then we should be able to adopt this system's conventions as to the delimitiation of graphical units. A possible solution would be to introduce a new interface requirement concerning the simple words' system: for a given sequence of characters (i.e. a MWU), identify all graphical units appearing in it. Then, the inflection graphs for MWUs would have to be compatible with the admitted delimitation conventions, as they are compatible with the morphological classes, categories and values defined for the simple words.

Another possibility would be 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.

8. 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 compact representation of numerous inflected forms due to a unification mechanism. It also allows for a joint description of different types of variation.

9. Acknowledgements

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

10. References

Anscombre, J.-Cl., 1990. Pourquoi un moulin à vent n'est pas un ventilateur. *Langue Française*, 86.

Arampatzis, A. T., T. Tsoris, C.H.A. Koster, and Th.P. van der Weide, 1998. Phrase-based information retrieval. *Information Processing & Managment*, 34(6):693–707.

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

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.

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.

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.

Przepiórkowski, A. and M. Woliński, 2003. The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish. In *The Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*.

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., 1993. Les groupes nominaux productifs et les noms composés lexicalisés. *LingvisticœInvestigationes*, 17(2).

Silberztein, M., 2003. *Dictionnaires électroniques et analyse automatique de textes : Le système INTEX*. Paris: Masson.

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