# Morphosyntactic Lexica in the OAL[1] Framework: Towards a Formalism to Handle Spelling Variants, Compounds and Multi-word Units

Helena Blancafort, Syllabs; Universitat Pompeu Fabra, Barcelona, Spain
Javier Couto, Syllabs, MoDyCo, CNRS, Université Paris X, France
Somara Seng, Syllabs

*The creation and maintenance of lexicographic resources are labour-intensive tasks. In this paper we present SylLex, a formalism to encode morphosyntactic lexica and how it is used in the OAL framework, a tool to aid the linguist to create and maintain such resources in an industrial context. The aim was to have a user-friendly and ready-to-use tool as well as an intuitive and easy-to-use formalism, SylLex, so that a linguist without previous experience can work effectively after one or two hours of using the tool. In this paper, we describe the formalism SylLex and explain how variants, compound words and multi-word units are encoded by SylLex and managed with OAL. Finally, we discuss the advantages and disadvantages of the formalism and OAL tool, and mention further work.*

## 1. Introduction

Linguistic resources are at the core of most NLP applications. A common low-level analysis necessary to many NLP applications is part of speech (POS) tagging that uses morphosyntactic lexicons to assign to each token (word in a corpus) morphosyntactic information and then selects the right information out of the possible POS tags and lemmata. There are different types of POS tagging algorithms using more or less linguistic information. However all of them need a morphosyntactic lexicon and the quality of the output depends on the quality of the lexicon: missing entries in the lexicon will have to be guessed by the POS-tagger, errors in lexicon entries have a direct impact on the POS output.

Still, in industrial NLP contexts, the development of morphosyntactic lexica in the NLP domain is too often done manually, without specific tools, and computational linguists have to handle all the data with scripts and/or standard text editors. Thus, the information is often stored in textual format. To improve and provide quality assurance of the development of linguistic resources, Couto et al. (2010) have designed and implemented OAL, an architecture to provide a user-friendly environment to develop lexicographic resources. Fontenelle (2008), Joffe and Schryver (2004), and ten Hacken (2002) and ten Hacken et al. (1994) present further tools for computer-assisted lexicon development. The novelty of OAL is to assist the linguist during the whole process, from corpora compilation to quality assurance of the resources (for more details about the software's architecture see Couto et al. (2010)). Moreover, it provides a user-friendly and ready-to-use tool as well as an intuitive and easy-to-use formalism, SylLex, so that a linguist without previous experience can work effectively after one or two hours of training.

The SylLex formalism implemented in OAL has been designed to represent morphosyntactic lexica. The idea behind was to have a tool to enhance the maintenance of the lexicon, this is, to improve the access to lexical data and accomplish lexicon maintenance task in a more efficient way, assuring tags consistency, lexicon entries correctness, and avoiding the generation of further errors when adding new forms. As the lexicon is organised by group of lemmas sharing a same inflection paradigm, we designed an interface that presents the lexicon data ordered by lemmas and their corresponding inflection patterns. Inflection rules are written in a traditional way, this is, by filling inflection patterns in an interface, as we will see in the next section. The linguist just needs some basic knowledge of regular expressions and does not need to learn a complicated formalism to encode inflection rules; the aim was to perform lexicographic work in a way as intuitive as possible.

[1] OAL stands for French 'Outil d'aide au linguiste', this is, a tool to assist the linguist.

An interesting tool is Word Manager (WM), a tool to develop dictionaries that handles inflection as well as word formation (ten Hacken 2002; ten Hacken et al. 1994). As opposed to OAL tool, linguistic and lexicographic knowledge are separated which means that two different profiles, a linguist and a lexicographer, do separate work with separated interfaces. The linguist encodes the morphological rule system of the language while the lexicographer has a specific interface to introduce new words that s(he) can link to existing inflection rules or to word formation rules to generate new words. In OAL there is no distinction between linguistic and lexicographic work, as the formalism to encode the rules is very simple and intuitive. A new user can rapidly and easily understand inflection paradigms and write new inflection rules. Real-life experience has shown that few hours suffice for a new linguist to encode new words and be able to handle paradigms (creation and modification). However, in contrast to WM, OAL does not include word formation rules. The focus of OAL relies on the whole procedure of lexicon creation and maintenance, as on the integration of a guesser of inflection paradigms to help the linguist by suggesting possible inflection paradigms for a new word found in a corpus.

In addition to this, the formalism includes the possibility to encode different spelling variants of a form. Another important issue is the handling of compound and multi-word units. The same formalism is used for lexica in several languages: French, Italian, Spanish and English for the time being; German, Polish and Chinese are to be included this year.

In this paper, first we present the SylLex formalism. Then we discuss the issue of variants and in section four we describe the solution adopted to handle all type of compounds as well as multi-words. Finally, we draw conclusions on the advantages and disadvantages of the formalism and OAL and mention further work.

## 2. SylLex: a formalism to encode morphosyntactic lexica

SylLex is the formalism of our morphosyntactic lexicons and organizes every lexicon in three components: a list of lemmas, a set of inflection paradigms and a set of patterns. Patterns help the linguist to create inflection paradigms: they contain the information about all morphosyntactic tags, all possible inflection forms of type of lemma and further morphological information (e.g., stems and suffix) needed for word inflection that the linguist will specify when creating an inflection paradigm. In French, for instance, we have a single pattern for nouns, two patterns for adjectives (qualifiers and ordinals), and two patterns for verb inflection: auxiliary and full verbs, as the part-of-speech tag is different for each of them.

So when a new lexicon is created, first the patterns are outlined and then inflection paradigms are written for a first lemma that is used as a prototype. Finally, a list of lemmas sharing the same inflection paradigm is associated to the resulting inflection paradigm. As a result, the lexicon is organized by lemmas sharing a same inflection paradigm.

### 2.1. Lemmas and Inflection Paradigms

As explained in Loupy and Gonçalves (2008), there are different ways of organizing those lexicons, as the common *form-lemma-tag* format that associates to each form a lemma and its corresponding tag, or a similar format that to each form associates all possible lemmas and tags as in Freeling's dictionaries (Carreras et al. 2004) to explicit the ambiguity of a single form. Loupy and Gonçalves (2008) use a formalism that organizes the lexicon in two components: the first one lists all the lemmas with an inflection class associated to each one, and a second component with all the corresponding inflection paradigms, similar as done in DELAS (Gálvez 2006) and the French Lefff lexicon (Sagot et al. 2006) and its Spanish equivalent Leffe (Molinero et al. 2009). The advantage of doing so is to reduce the size of the

lexicon and to handle the lexicon more easily. When an error concerning an inflection paradigm is corrected, the correction inside the inflection paradigm suffices; it will be applied to all corresponding lemmas, whereas when using an extensive format as the form-lemma-tag one, each form concerned has to be corrected. This is usually done with a text editor or a script and sometimes generates further errors in the lexicon.

### 2.2. Naming convention of paradigms

The other formalisms presented above, like the Leffe formalism (Molinero et al. 2009), show two disadvantages: first, the name of the inflection classes is just an alphanumeric code and does explicit further information, as *V4* in the example below for a specific subgroup of Spanish verbs on suffix *–ar*. Secondly, paradigms are just listed and the linguistic relationship between morphologically related paradigms is ignored. Figure 1 shows the lexicon entry for a lexicon organized in lemmas (figure 1).

```
destacar V4 Lemma;v; <arg0:Suj:cln|scompl|sinf|sn,arg1:Obj:cla|scompl|sn>; %actif, %passif, %ppp employ´e comme adj
```

Figure 1. Inflection rule for Spanish verb lemma *destacar* in Leffe's lexicon organized in lemmas

Our naming convention first expresses the part-of-speech. For nouns and adjectives, gender information is indicated to know whether masculine (Masc), feminine (Fem), or both genders (MascFem) are to be generated. Then we indicate the final suffix for each inflection: the one for the singular form and the one for the plural form: *N_Masc/al-aux* stands for French masculine names with singular suffix *–al* and plural suffix *–aux*. Adjectives follow the same convention. Furthermore, for each paradigm a prototype is shown to enrich the information. For verbs, we first indicate the name of the group (first, second or third group following French grammar's tradition), the ending suffix and then information about the past participle, whether it is invariant or no, as *V1/er_ppFLEX*, for prototype *chanter* and *V1/ger_ppFLEX*, for prototype *manger*. Nooj's dictionaries (Siberzstein, 2005) use a naming convention as well, but just give the name of a prototype, no further information is given. Moreno-Sandoval and Goñi-Menoyo (2002) also just give the name of a prototype that they call *model*.

For completely irregular lemmas or unproductive rules applied to less than four lemmas with a specific inflection, we just indicate the name of the lemma, as done for French verbal lemma *aller*.

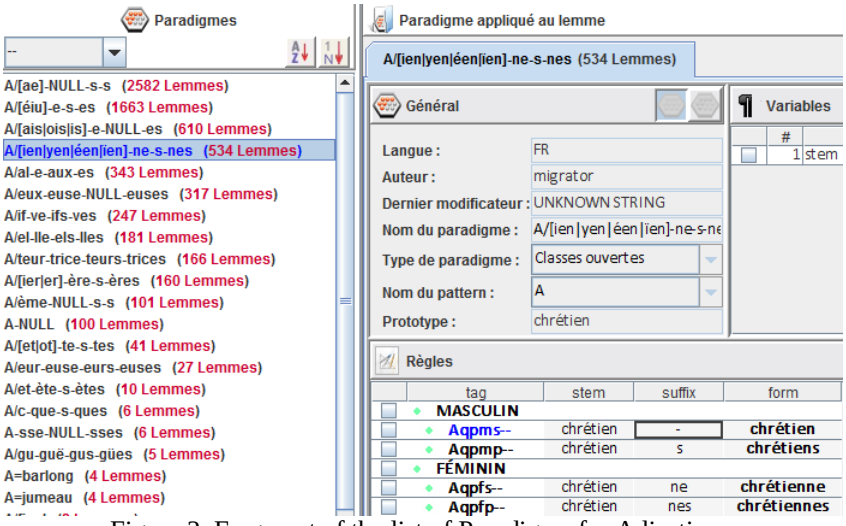


Figure 2. Fragment of the list of Paradigms for Adjectives.

Other lexica such as the Lefff separate exceptions and store the lemmas with a special inflection paradigm in a special file. As SylLex is integrated in a tool for lexicon development

and as a matter of internal coherence, we apply the same formalism even for non productive paradigms. As a consequence, the list of inflection paradigms is very long, but the OAL lexicon editor allows us to filter paradigms by frequency of application (see figure 2).

OAL aims at enhancing lexicon maintenance and quality assurance. The linguist can indeed see the productivity of each inflection paradigm and check how often it is applied to a lemma to be conscious of the impact that will have any modification carried out on a single paradigm.

### 2.3. Function words

Nevertheless, the formalism and especially the naming convention presented here is more difficult to handle for function words, where inflection paradigms are not productive and thus, the list of inflection paradigms is extremely long. Today OAL contains 337 inflection paradigms for 416 French function words, 288 paradigms for 518 Italian function words, and 198 for 603 English function words. Words that are not inflected at all and share the same morphological tag are regrouped together. For instance, all conjunctions of coordination are stored under a single paradigm *Coord_conj*.

### 2.4. Stem and suffixes

It is worth mentioning that the notion of stems and suffixes slightly differs from the ones used in linguistics. The suffix corresponds to the lexical string to be added to a stem to get a form. To define the stem, the linguist uses regular expressions as illustrated below (cf. figure 3): the regular expression between the two first slashes indicates the string to be matched.

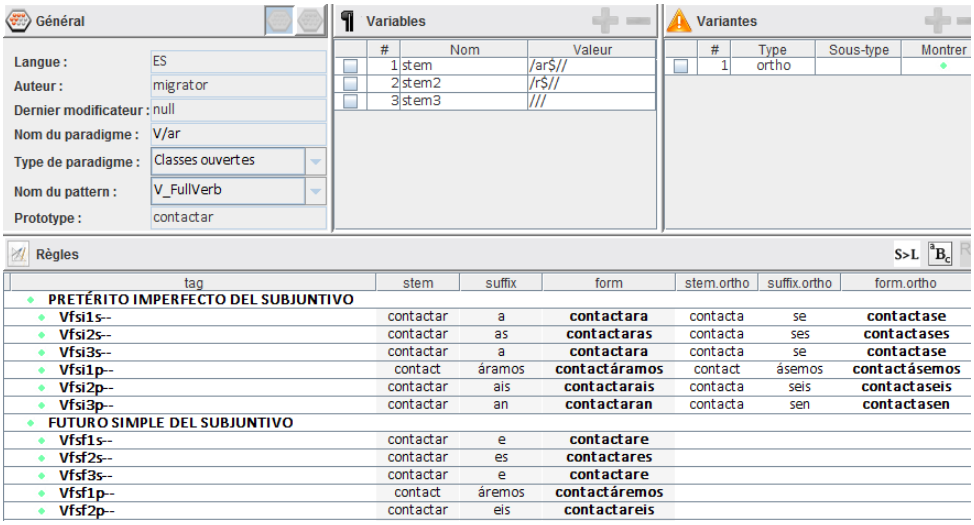

Figure 3. An inflection paradigm example.

The information that follows indicates the lexical string that will be replaced. As a result we obtain the stem to be used for one or several forms. For instance, to conjugate the Spanish regular verb *contactar*, there are three stems: *contact*, *contacta* and *contactar*, the lemma itself, without any modification.

### 2.5. Linking Morphologically Related Paradigms: the Notion of Inheritance between Paradigms

As far as we know, the formalisms presented above that organise the lexicon by lemmas do not express any link between paradigms and related paradigms, nor do they apply any patterns to enhance and control the creation of new inflection paradigms. In Romance languages, we find several productive paradigms for irregular forms that just differ in some inflection forms and that are interesting to maintain. In French, for instance, the most productive verbal

paradigm corresponds to verbs ending in *–er* (called *verbs of the first group*) as *chanter*. There are several variants of this paradigm for other verbs with the same ending but showing several irregularities, such as the verbs ending in *-ger* that have specific inflected forms following morphotactic rules. To resolve these problems, inflection paradigms were renamed and the notion of inheritance was added to link related paradigms. With SylLex, the paradigm of the verb *manger* inherits from the paradigm for the verbs of the first group. We still have two different paradigms; but the relation of inheritance is explicit and the linguist only has to encode the rules that are different from the 'father' paradigm. This notion of inheritance has already been used in well-known formalisms as DATR (Evans and Gazdar 1990). DATR represents an interesting alternative to finite-state-automatas (Karttunen 1992) to handle languages with rich inflection as Romance languages.

The theoretical morphological framework behind DATR is the Network morphology (Corbett and Fraser 1993) that already makes use of inheritance and overwriting to describe morphology. With the DATR formalism it is possible to have a paradigmatic approach as described in the Word-and-Paradigm model (WP) (Carstairs 1987). This approach is especially interesting for languages with rich inflection governed by paradigms and rules and presenting suppletion and defective forms as Spanish, see Moreno-Sandoval and Goñi-Menoyo (2002). For these languages the finite-state-morphology (Karttunen et al. 1992) and the two-level-morphology (Koskenniemi, 1983) based on phonological and spelling changes is less suitable than for agglutinative languages as Finnish. A paradigmatic approach suits bests for most phenomena. The drawbacks of the paradigmatic model are that it can be redundant and expensive for regular phenomena such as phonologically conditioned allomorphy or morpheme agglutination, for which a phonological approach is more suitable. A paradigmatic approach based on inheritance can reduce redundancy as shown for Spanish inflection in Moreno-Sandoval and Goñi-Menoyo (2002) by linking shared inflection processes. These authors have studied all the possible inheritance relations and established several hierarchal link types as stem allomorphy, suffix allomorphy and conjugation type. It is an exhaustive and rich study, but quite complex, especially for verbal inflection. In SylLex, we just apply inheritance for verbs of a same group, as the verbs in *ar* or *ir* in Spanish or first group and second group in French. Again the idea behind SylLex and OAL is that the linguist can work effectively in a few hours. As opposed to DATR, that aims at being a programming language for morphology rules, SylLex formalism aims at simplicity and has been defined to be integrated in an industrial easy-to-handle tool. As a consequence, while in DATR the user has to explicitly manipulate nodes and write several rules to encode inheritance relations, in SylLex the user only has to indicate the father paradigm and click on the entries that are different and do not inherit the inflection rules by redefining the stem and the suffix.

## 3. Variants

The problem of dealing with variants of a lemma or a form is a common concern in lexicon maintenance. A word can have different variants according to several spelling conventions depending on the country (geographical variants such as *color* in American English and *colour* in British English), on *real* use (a canonical normalized form and the variants used) or diachronically motivated (*clé* and *clef* in French). Other variants concern the register, such as informal variants of German verb forms in the present where the last vowel is omitted as in *hab'* or *hab* instead of *habe*. Moreover, in Spanish, verbs have systematically two different forms at the imperfect tense of the subjunctive mood as *cantase* or *cantara*. In figure 3 we can see how the variants are visualized in the OAL lexicon editor.

Some variants are generated systematically, this means that the phenomenon is recurrent and thus, that the lexicon can generate those forms automatically. Take for instance the forms with

β in German. In Swiss German, the German character β is not used and is systematically replaced by double s, as in *Fuss* instead of *Fuß*. This is why a further function of OAL is to generate those forms automatically. With OAL those variants can be encoded as a variant, which means that we explicit that the form is not exactly an inflected form but a variant of an inflected form. This information is crucial for search applications.

This is also necessary in English to handle the problem of variants of compounds, where the fact that a compound is written in separate tokens or one single token or lexeme (joint) depends on the level of lexicalization of the word. Thus, a less lexicalized word such as *tagset* should have a variant *tag set*. In French, some hyphenated compounds can be spelled without the hyphen as well, as *porte-feuille* and *portefeuille*. Thus, we need to handle the variants of compounds and multi-words. Next, we will explain how we handle compounds in OAL.

## 4. Handling of compounds and multi-word units

In morphology, the definition of multi-word units (MWUs) and compounds is quite controversial (Habert and Jacquemin 1993; Corbin 1992), as even the basic concepts used to define them are controversial. The same applies for the measures that can be used to define a compound, as the degree of non-compositionality. Different authors as Bauer (2003), Downing (1977) and Booij (2005) give different definitions for MWE but agree in the fact that MWUs are formed by two or more words, share a degree of morphological, distributional and semantic non-compositionality, and have unique and constant reference as mentioned in Savary (2005). She gives a pragmatical definition of MWUs: linguistic objects placed on the frontier between morphology and syntax, and that form a contiguous sequence of *graphical units* which depending on the application are to be listed in the lexicon and process as a unit.

In the scope of our work, we understand compounds as a combination of lexemes that form a single and larger word (Booij 2005 and Bauer 2003), while we define MWU as a word formed by two or more words separated by a blank. Compounds can be agglutinative (stems joined together without any marker as *sunflower*) or hyphenated (*week-end*). The elements of MWUs are written separately and may include a preposition or not (French *pomme banane* or *pomme de terre*). When developing morphosyntactic lexica, several questions arise concerning MWU: the degree of lexicalization of a MWU to decide its inclusion in the lexicon or not, and secondly, the identification of the head (Booij 2005). One of the tests to decide the inclusion of a MWU in the lexicon include semantic criteria as the notion of transparency versus opaque word formation, the latter needed to be included in the lexicon as the meaning is not easily deductible from its components, as in *Morgenrock* in German or *cupboard* in English. A further delicate issue is the identification of the head of the compound or multiword that is necessary to assign morphosyntactic information such as gender, number as well as the inflection paradigm. Compounds with a head assigning morphosyntactic information are called *endocentric* compounds. Other compounds are *exocentric* and as such, do not have an obvious head providing categorical information. The head is outside the construction itself (Bauer 2009). This is the case of noun compounds consisting of a verbal stem with a plural noun as *lava-piatti* and *limpia-botas* where the number nor the gender of the constituent provides the final compound with morphological information.

Regarding the position of the head, in Germanic languages the right constituent of the compound is normally the head (Booij 2005). Booij provides examples of left-headed and right-headed compounds in a same language as in Italian, with left-headed compound *capostazione* and right-headed compound *croce-rossa*. Booij discusses the issue of lexicalized phrases and word-hood: what is the frontier between compounds and phrases? One possible interpretation is that as *capostazione* inflects word-internally to *capistazione*, one could

consider that it is not a word, but a phrase and consider Italian with right-headed compounds only. The boundary between compounds and phrasal lexical expressions is not always clear and tackles the issue of the relation between morphology and syntax. There are different criteria to define compoundhood, as stress and phonological criteria, syntactic criteria as impenetrability and inseparability, and inalterability and the behaviour of the complex item with respect to inflection (Lieber and Štekauer 2009). Spelling is often rejected, due to the inconsistency of compounds spelling in languages like English and French (*blackboard* vs. *black-board*, *porte-feuille* vs. *porte feuille*).

As we have seen above, the position and identification of the head of a compound is quiet complex. To handle compounds and MWUs, instead of writing specific inflection paradigms for each case, we decided to adopt a more practical solution introducing the notion of pivot. The pivot is the element that contains the morphosyntactic information. If the compound does not have an obvious head as *porte-monnaie*, we apply to the whole word the inflection rule corresponding to a masculine noun with plural form ending in *s*. Here there is no pivot, the compound does not inherit the morphological information of the lemma *monnaie*, but it takes the information of the inflection paradigm associated to the compound *porte-monnaie*. Next, we will illustrate some examples with figures of OAL.

The first one (figure 4) is an example of the pivot being the head of the compound, both elements are inflected. We can see that the right-element *bateau* is designed as a pivot and thus, gives the morphosyntactic information (masculine noun). For the second element, *mouche*, the same inflection rule as for the already existing lemma *mouche* applies. Both elements do agree in number and are inflected following the inflection paradigms indicated at the right window on the top.

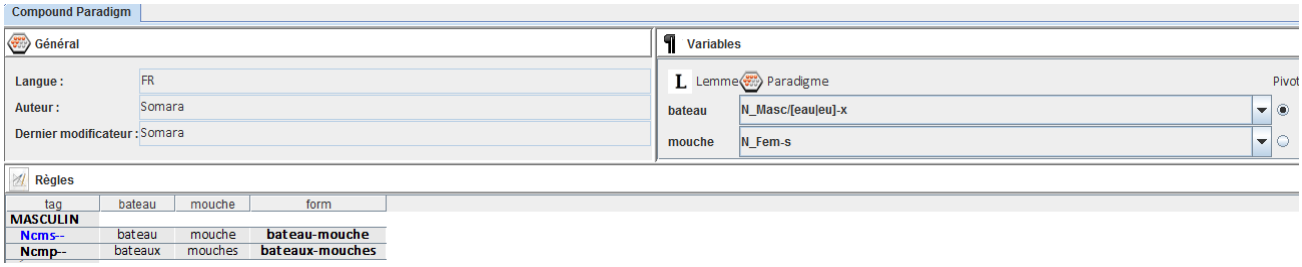


Figure 4. Paradigm for French compound *bateau-mouche*

The same applies for MWU where both elements inflect, as in *cousin germain*.

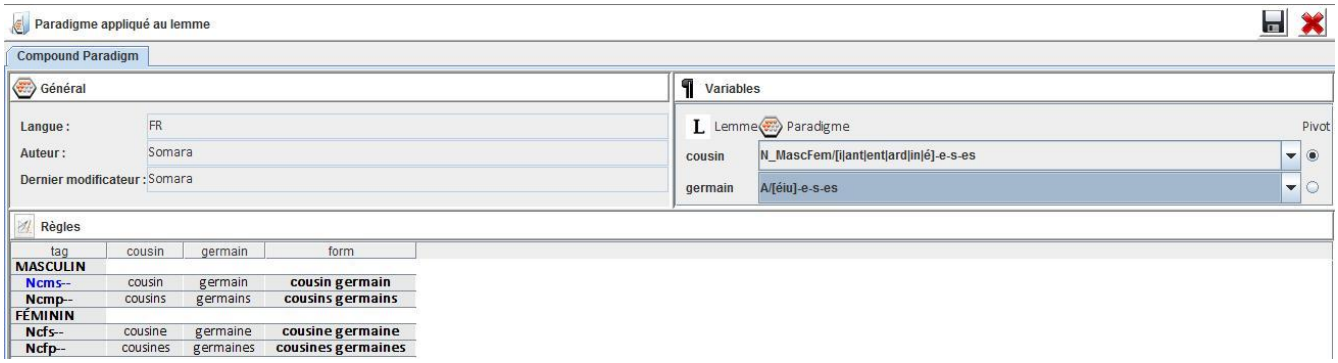


Figure 5. Paradigm for French MWE *cousin germain*

Lexicographic tasks performed with SylLex formalism and its implementation in the OAL tool seem to us more user-friendly and easier to learn for a new user than other systems like Multiflex using graphs and presented in Savary (2005, 2008). Below we find an illustration of a graph used to inflect *bateau mouche* as a MWU with Multiflex. In our opinion, writing

those graphs is quite time-consuming and do not take advantage from already existing information about inflection for the elements of a MWU.

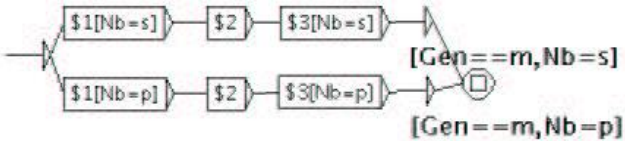


Figure 6. Inflection of *bateau mouche* with Multiflex (Savary, 2005, 2008)

The figure below illustrates a compound where only the head is inflected, while the modifier is not inflected at all. It suffices to indicate to the system that the element is not the pivot and that it does not inflect.

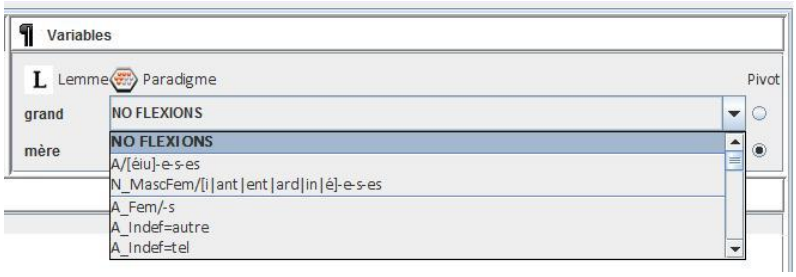


Figure 7. Compound word where only the head is inflected

One of the problems encountered with the inflection of MWU was the case when the modifier was not the lemma, but a form, which means that only some inflection forms are used, as mentioned and illustrated in Savary (2008). The MWU *mémoire vive*, for instance, modifier is inflected, but is not a lemma. *Vive* is the feminine form of *vif*. To resolve this problem, we select a different rule, as if *vive* was a feminine lemma adjective, as illustrated below.


Règles

| tag | mémoire | vive | form |
|---|---|---|---|
| *MASCULIN* | | | |
| *Ncms--* | | | |
| *Ncmp--* | | | |
| **FÉMININ** | | | |
| **Ncfs--** | mémoire | vive | **mémoire vive** |
| **Ncfp--** | mémoires | vives | **mémoires vives** |

Figures 8 and 9. Compound word where all constituents are inflected and the resulting forms

Savary (2008) discusses this problem and explains that the DELAC dictionary (Silberztein, 1993) introduces the notion of *artificial lemma* to solve this problem. With SylLex, we do not generate artificial lemmas, but we do not link the lemma *vif* to the MWU *mémoire vive*.

## 5. Conclusion and Further Work

In this paper we have presented the formalism SylLex used to encode morphosyntactic lexica in the OAL framework that aids the linguist to develop linguistic resources in an industrial context. By dividing the lexicon in three components, lemma, inflection paradigms and patterns, the linguist can better manipulate the data, assure consistency, and maintain the lexicon by modifying inflection paradigms instead of dealing with scripts to directly modify lexicon entries in text files. In addition to this, the system takes advantage of the notion of inheritance between different inflection paradigms. Moreover, we saw that naming conventions are important to get a better overview of all the inflection paradigms. We further evoked the problem of spelling variants and the complexity to handle compound and MWUs. With OAL we do not write specific rules for the inflection of compounds and MWUs; we reuse information already encoded in the single elements of the MWUs that is already stored

in the lexicon entries of simple nouns or adjectives.

The main feature of our approach is the integration of a powerful and flexible formalism in a user-friendly and efficient lexicon editor that allows the linguist to deal with linguistic notions such as inflection paradigms, lemmas and suffixes instead of handling a huge file in a text editor that simply lists all the forms of a lexicon. Other formalisms, such as DATR or Leffe are, in our opinion, more complex and hard to learn. Thus, their application in an industrial context is less appropriate. OAL provides a more user-friendly and intuitive environment for lexicon development, speeds up the whole process, and uses a formalism, SylLex, that is also intuitive and easy to understand. Moreover, OAL does not only provide a lexicon editor, but different functionalities to assist the linguist during the whole process of lexicon enrichment: it includes a guessing tool to aid the linguist to add new words to the linguist by suggesting possible inflection paradigms for each new lemma for simple nouns, adjectives and verbs. Further work will concentrate on how to aid the linguist to encode compounds and MWUs, which will be very helpful for encoding terminologies with specific compounds and MWUs. The next challenge for OAL will be to add other languages with rich inflection as German and Polish, as well as an analytic language, Chinese.

## References


Bauer, L. (2009). 'Typology of compounds'. In Lieber, R.; Štekauer, P. (eds). *The Oxford Handbook of Compounding*. Oxford: Oxford University Press. 343-356.

Bauer, L. (2003). *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Booij, G. (2005). *The grammar of words. An introduction to linguistic morphology*. New York: Oxford University Press.

Carreras, X.; Chao, I.; Padró, L.; Padró, M. (2004). 'FreeLing: An Open-Source Suite of Language Analyzers'. In *Proceedings of LREC'04*. Lisbon.

Carstairs, A.D. (1987). *Allomorphy in Inflexion*. London: Croom Helm.

Corbett, G.G.; Fraser, N.M. (1993). 'Network morphology: A DATR account of Russian inflectional morphology.' In *Journal of Linguistics*.

Corbin, D. (1992). 'Hypothèses sur les frontières de la composition nominale'. In *Cahiers de grammaire* 17.

Couto, J.; Blancafort, H.; Seng, S.; Talby, A.; Loupy, C. de (2010). 'OAL : A NLP Architecture to Improve the Development of Linguistic Resources for NLP'. In *LREC 2010* Malta.

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

Evans, R.; Gazdar, G. (1990). *The DATR Papers*. Brighton: University of Sussex.

Fontenelle, T.; Cipollone, N.; Daniels, M. ; Johnson, I. (2008). 'Lexicon Creator: A Tool for Building Lexicons for Proofing Tools and Search Technologies'. In *Proceedings of EURALEX 2008, Barcelona, Spain.*

Galvez, C. (2006). 'El diccionario electrónico: un instrumento para la unificación de términos en la indización automática'. In *Linguax: Revista de Lenguas Aplicadas.*

Habert, B. ; Jacquemin, C. (1993). 'Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques'. In *Traitement Automatique des Langues* 2.

Joffe, D.; Schryver, G-M. (2004). 'TshwaneLex – A State-of-the-Art Dictionary Compilation Program'. In *Proceedings of EURALEX 2004.* Lorient.

Karttunen, L.; Kaplan, R. M.; Zaenen, A. (1992). 'Two-level Morphology with Composition'. In *Proceedings of Coling 92.* Nantes.

Koskenniemi, K. (1983). *Two-level Morphology. A General Computational Model for Word-Form Recognition and Production.* Department of General Linguistics, University of Helsinki.

Lieber, R.; Štekauer, P. (2009). 'Introduction: Status and Definition of Compounding'. In Lieber, R.; Štekauer, P. (eds). *The Oxford Handbook of Compounding*. Oxford: Oxford University Press. 3-18.

Loupy, C. de; Gonçalves, S. (2008). 'Aide à la construction de lexiques morphosyntaxiques'. In *Proceedings of EURALEX 2008.* Barcelona.

Moreno-Sandoval, A.; Goñi, J.M. (2002). 'Spanish Inflectional Morphology in DATR'. In *Journal of Logic, Language and Information*.

Molinero Miguel A.; Sagot, B.; Lionel, N. (2009). 'Construcción y extensión de un léxico morfológico y sintáctico para el Español: el Leffe'. In *SEPLN 2009.* Donostia.

Sagot, B.; Lionel, C. ; Clergerie, É. de la; Pierre, B. (2006). 'The Lefff 2 syntactic lexicon for French: architecture, acquisition, use'. In *Proceedings of LREC 06.* Genova.

Savary, A. (2005). 'Towards a Formalism for the Computational Morphology of Multi-Word Units'. In Vetulani (ed.) *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of LTC 2005.* Poznań.

Savary, A. (2008). 'Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches'. In *Linguistic Issues in Language Technology* 1 (2). 1-53.

Silberztein, M. (1993). 'Les groupes nominaux productifs et les noms composés lexicalisés'. In *Lingvisticæ Investigationes* 17 (2).

ten Hacken, P.; Bopp, S.; Domenig, M.; Holz, D.; Hsiung, A.; Pedrazzini, S. (1994). 'A Knowledge Acquisition and Management System for Morphological Dictionaries'. In *Proceedings of Coling 94.* Kyoto.

ten Hacken, P. (2002). 'Word Formation and the Validation of Lexical Resources'. In González Rodríguez, M.; Paz Suárez Araujo, C. (eds.). *Proceedings LREC 2002.* Las Palmas.