Multiflex: a Multilingual Finite-State Tool for Multi-Word Units

Agata Savary*

Université François Rabelais Tours, France,
agata.savary@univ-tours.fr

Abstract. Multi-word units are linguistic objects whose idiosyncrasy calls for a lexicalized approach allowing to render their orthographic, inflectional and syntactic flexibility. *Multiflex* is a graph-based formalism answering this need by conflation of different surface realizations of the same underlying concept. Its implementation relies on a finite-state machinery with unification. It can be applied to the creation of linguistic resources for a high-quality natural language processing tasks.

Key words: multi-word units, finite-state morphology, Multiflex

Describing the variability of multi-word units Multi-word units (MWUs) encompass a number of hard-to-define linguistic objects: compounds, complex terms, named entities, etc. They are composed of two or more words, and show an important degree of flexibility on different levels: orthographic (*head word* vs. *headword*), inflectional (*man servant* vs. *men servants*), syntactic (*birth date* vs. *date of birth*), and semantic (*hereditary disease* vs. *genetic disease*). This flexibility is hard to represent precisely and exhaustively within general grammar-based models due to idiosyncrasy (e.g. *chief justices* vs. *lords justice*).

Multiflex is a formalism and a tool that copes with flexibility and idiosyncrasy of MWUs by a fully lexicalized two-layer approach. Figure 1 shows the description of a German MWU whose inflection and variation paradigm is given in examples (1) and (2). The sequence is segmented into tokens (here \$1 through \$7) by the underlying module handling the morphology of single words. The possibly inflected tokens are annotated by their lemmas, morphological features, and any data needed to generate other inflected forms of the same unit.

(1) *Organisation der Vereinten Nationen* :neF:aeF:deF:geF
'United Nations Organisation' in singular (e) feminine (F) nominative (n), accusative (a), dative (d) and genitive (g)

(2) *Vereinte Nationen* :nmF:amF; *Vereinten Nationen* :nmF:amF:dmF:gmF
'United Nations' in plural (m) with a determined or undetermined adjective

A path in a graph starts with the leftmost edge and ends with the final encircled box. The morphological information contained in the boxes refers to

* The project is partially financed by the Polish Ministry of Science and Higher Education, decision number 567/6. PR UE/2008/7.

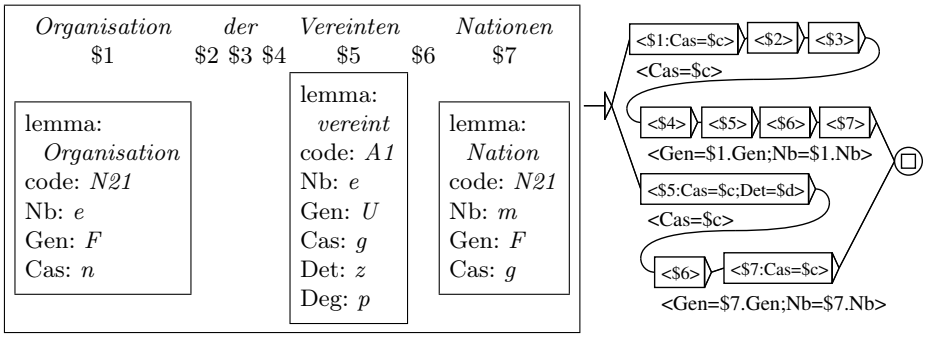

Fig. 1. Lemma annotation and inflection graph of a German MWU

the constituents of the MWU while the one placed under a box refers to the morphological description of the resulting MWU inflected form. Both types of information usually take the form of possibly uninstantiated and partial feature structures. Here, the upper path describes all forms in example (1). Constituents \$2 through \$7 are recopied as such from the MWU lemma, while constituent \$1 (*Organisation*) is inflected for any case due to the unification variable $\$c$ which can take any value from the case domain in German. The lower path represents all elliptic variants in example (2). Constituents \$1 through \$4 are omitted while constituent \$7 (*Nationen*) shifts to the head position and becomes case-inflected. The modifier \$5 agrees with the new headword in case (same unification variable $\$c$) and inflects for determinedness. In the full form (upper path) the morphological features of the whole MWU are inherited from the first constituent. The number and the gender are those that \$1 takes in the MWU lemma ($\langle Gen = \$1.Gen; Nb = \$1.Nb \rangle$), here *eF*, while the case is as in the particular MWU inflected form ($\langle Cas = \$c \rangle$). In the elliptic form (lower path) the same kind of inheritance occurs with respect to the seventh constituent.

The use of unification variables allows for a compact description of unification paradigms. Here, the 10 forms would need 10 different paths if no unification variables were available. In highly inflected languages, such as Slavic languages, this facility is crucial: although many compounds may have several dozens of forms, a unique path is often enough to render them all.

Finite-State Machinery *Multiflex* is inspired by the Paris school of finite-state morphology. It uses the graph editor of the *Unitex* system [9], and its generic finite-state library for binary representation and exploration of graphs (boxes and arrows in graphs correspond to transitions and states in finite-state transducers). However the semantics introduced in *Multiflex*' graphs is novel, although formally close to decorated RTNs in [3], regular expressions with feature structures in [4], and flag diacritics [2]. It represents a meta-grammar: (i) each compound with its tokenization and annotation is a rule, (ii) each inflectional graph is a meta-rule, i.e. the transformations that can be applied to a rule in order

to produce new rules (compound inflected forms). This view is inspired from [5]. However in *Multiflex* all transformations (except in embedding) are gathered within the same metarule. Thus the dependencies between meta-rules are very scarce, which avoids the problem of a "card tower" in traditional grammars (the modification of a rule perishes the validity of other rules). This highly modular aspect of *Multiflex* rules makes their management and debugging easier.

At present, *Multiflex* operates in the generation mode. When applied to an annotated compound it performs the depth-first search exploration of the minimal finite-state transducer behind the graph. A transition is followed if its input and its output labels are sound. An input is sound in one of the two cases: (i) it is a constant string, (ii) it refers to an existing component (*$8* would be invalid in Fig. 1) and all the category-value equations (if any) can be fulfilled. The last condition means that: (i) categories are relevant to the component (unlike *Det* for $7 in Fig. 1), (ii) values belong to the domains of their categories (*Nb=masc* is incorrect), (iii) unification, if any, can be performed. If a unification variable has already been instantiated in a previous transition on the same path then its value must belong to the right category, and it must be accepted by the inflection paradigm of the current constituent. If however a unification variable has not yet been instantiated, it is instantiated to each value of its category's domain for each outgoing path. Thus, each path represents at least as many forms as there are allowed combinations of all unification variables it contains. An output label is sound if the category-value equations can be fulfilled: (i) the values belong to their categories' domains, (ii) if a value is fixed, its category has not yet been associated with a different value, (iii) if the value is inherited it refers to an existing component and a relevant category, (iv) unification can be performed.

While exploring a graph, *Multiflex* collaborates with an external morphological module for single words. This module must share the same morphological model (up to identifier replacement), must provide a clear-cut definition of a token boundary, and must generate on demand particular inflected forms for single tokens. Its implementation is not necessarily based on finite-state machines. *Multiflex* has been successfully interfaced with two underlying modules, one FSM-based ([9]), and one using a relational database ([14]).

Applications and Evaluation Our first motivation for an inflection tool for MWUs came from the FSM-toolkit *Intex* [12], and led to a prototype which was applied to the creation of two DELA-type electronic lexicons of (general and terminological) English compounds (about 60,000 lemmas and 110,000 inflected forms each). The first one is distributed with *Intex* and *Unitex*, the second one was used in a translation aid software *LexProCD Databank* for term extraction.

Later our formalism was improved and re-implemented as *Multiflex*. It was released with *Unitex* (under the LGPL license), where it is used for an automatic generation of electronic lexicon of compound inflected forms (the so-called DELACF) which are matched against a corpus during the process of morphological analysis. It was tested on a 2000-entry sample of a Serbian MWU lexicon [8], and on examples of French, German, Polish, Portuguese and English. *Mul-*

tiflex is also a part of two encoding support tools: (i) *WS2LR* [7], which allows an automated controlled encoding of morphological dictionaries, aligned corpora and wordnets in Serbian, (ii) *Toposław* [11], an outcome of the European LUNA project (`http://ist-luna.eu`) supporting controlled description of Polish toponyms in written and spoken corpora. Finally, *Multiflex* is incorporated into to the linguistic interface of the multilingual ontology of proper names *Prolex* [13].

In [10] a large contrastive study of 11 lexical approaches to the inflection and variation of MWUs in 7 languages was performed. It analyzes a dozen linguistic properties of MWUs (exocentricity, irregular agreement, defective paradigms, variability, etc.), and desirable descriptive and computational facilities (unification, non-redundancy, encoding interface, etc.). In the light of this study *Multiflex* belongs to the most expressive and effective tools along with *lexc* [6], *FASTR* [5], and *HABIL* [1]. Its drawbacks include the lack of modeling of derivational and semantic variants, abbreviations, and dependencies existing between a MWU and neighboring external elements. In the long run *Multiflex* needs to be enlarged to non-contiguous MWUs such as verbal expressions, admitting insertions of free external tokens. We also wish to integrate machine learning tools allowing both to acquire new data from the corpora and to predict inflection graphs for them.

References

1. Alegria, I., et al.: Representation and Treatment of Multiword Expressions in Basque. In: ACL Workshop on Multiword Expressions. (2004)
2. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI (2003)
3. Blanc, O., Constant, M.: Lexicalization of Grammars with Parameterized Graphs. In: Proceedings of RANLP'05. (2005)
4. Drożyński, W., et al.: Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. Künstliche Intelligenz **1** (2004) 17–23
5. Jacquemin, C.: Spotting and Discovering Terms through Natural Language Processing. MIT Press (2001)
6. Karttunen, L., Kaplan, R.M., Zaenen, A.: Two-Level Morphology with Composition. In: Proceedings of COLING-92, Nantes. (1992) 141–148
7. Krstev, C., Stanković, R., Vitas, D., Obradović, I.: Workstation for Lexical Resources - WS4LR. In: Proceedings of LREC'06, Genoa, Italy. (2006)
8. Krstev, C., Vitas, D., Savary, A.: Prerequisites for a Comprehensive Dictionary of Serbian Compounds. LNCS **4139** (2006) 552–563
9. Paumier, S.: Unitex 2.1 User Manual. www-igm.univ-mlv.fr/~unitex (2008)
10. Savary, A.: Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. Linguistic Issues in Language Technology **1**(2) (2008) 1–53
11. Savary, A., Rabiega-Wiśniewska, J., Woliński, M.: Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. LNAI **5070** (to appear)
12. Silberztein, M.: Dictionnaires électroniques et analyse automatique de textes : Le système INTEX. Masson, Paris (1993)
13. Tran, M., Maurel, D.: Prolexbase: Un dictionnaire relatonnel multilingue de noms propres. Traitement Automatiques des Langues **47**(3) (2006) 115–139
14. Woliński, M.: Morfeusz – a Practical Tool for the Morphological Analysis of Polish. In: Proceedings of IIS:IIPWM'06, Springer (2006) 503–512