Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX

Elisabeth Breidt
Seminar für Sprachwissenschaft
Universität Tübingen
Wilhelmstr. 113
D-72074 Tübingen
Germany
breidt@sfs.nphil.uni-tuebingen.de

Frédérique Segond, Giuseppe Valetto
Rank Xerox Research Centre
6, chemin de Maupertuis
F-38240 Meylan
France
segond@grenoble.rxrc.xerox.com
valetto@mailer.cefriel.it

Abstract

Most multi-word lexemes (MWLs) allow certain types of variation. This has to be taken into account for their description and their recognition in texts. We suggest to describe their syntactic restrictions and their idiosyncratic peculiarities with local grammar rules, which at the same time allow to express in a general way regularities valid for a whole class of MWLs. The local grammars can be written in a very convenient and compact way as regular expressions in the formalism IDAREX which uses a two-level morphology. IDAREX allows to define various types of variables, and to mix canonical and inflected word forms in the regular expressions.[1]

1 Introduction

Most texts are rich in multi-word expressions that cannot be properly understood let alone be processed in an NLP system, if they are not recognized as complex lexical units. Such expressions which we call multi-word lexemes (MWL) range from idioms (***to rack one's brains over sth***), over phrasal verbs (***to come up with***), lexical and grammatical collocations (***to make love, with regard to*** resp.) to compounds (***on-line dictionary***).

While certain MWLs only occur in exactly one form, e.g. ***out of the blue*** or G:*um Haaresbreite* ('by a hair's breadth', lit. by hair's breadth), and can thus be easily recognised with simple pattern matching techniques, it is well-known (see e.g. Gross 1982, Brundage et al. 1992, Nunberg et al. 1994) that most MWLs cannot be treated like completely fixed patterns, since they may undergo some variation. However, only a ***subset*** of the variations allowed by general rules is valid: outside that subset, the expression loses its special, idiomatic meaning, either reverting to its literal meaning or losing any significance altogether. In certain cases, MWLs can even contradict normal syntactic rules, as with ***by and large***, or G:*von Haus aus* ('originally', lit. from house out), where general rules would require an article between the preposition and the noun.

The identification of MWLs is essential for any natural language processing based on lexical information, ranging from intelligent dictionary look-up over concordancing or indexing to machine translation. Therefore, the restricted lexical and syntactic variability of MWLs and their idiosyncratic peculiarities need to be expressed in the computational lexicon in order to be able to recognize the full range of their occurrences. We propose to use local grammars for this, written as a special type of regular expressions (REs) in the finite-state formalism IDAREX which makes use of a two-level morphological lexicon. So far, we have successfully applied this approach to approximately 15,000 English, French and German MWLs (see also Segond and Breidt 1995).

2 Restricted Variability of MWLs

Basically, we recognize four types of variability (see also Fleischer 1982, Brundage et al. 1992, Engelke 1994) that a description of MWLs, both for NLP and for human use, should cover. Though part of the variability of MWLs may follow from their semantic properties as argued in recent work (e.g. Nunberg et al. 1994), it is difficult to establish such a relationship on a large scale, and a lot of remaining idiosyncratic characteristics of individual MWLs need to be represented.

Morphological Variation: Particular words in the MWL may undergo certain inflections.

For instance, in G:***durchschlagender Erfolg*** ('sweeping success', lit. rubbing-off success), noun

[1] Part of this work was funded under LRE 62-080 by the EEC.

and adjective can vary in case and in number, and comparative and superlative form are possible for the adjective, whereas G:*grüne Welle* ('phased traffic lights', lit. green wave) may only vary in case, but not in number or adjective comparison without loosing its idiomatic meaning.

Lexical Variation: One or more words can be substituted by other terms without changing the overall meaning of the MWL.

For instance, in F:*perdre la tête* ('to lose one's mind', lit. to lose the head), the noun can be substituted by *la boule* (lit. ball, coll. head) or *les pédales* (lit. pedals) without loosing its idiomatic meaning, but not by *la tronche* (lit. slice, coll. head).

Modification: One of the MWL's constituents can be modified, preserving the idiomatic meaning.

For instance, in G:*den (schönen) Schein wahren* ('to keep up appearances', lit. the (nice) pretence preserve) the presence or absence of the adjective does not change the meaning at all, whereas in G:*das Handtuch werfen* ('to throw in the towel', lit. the towel throw) any modification would evoke the literal meaning.

Structural Flexibility: This includes phenomena like passivization, topicalization, scrambling, raising constructions etc.

For instance, whereas in German standard word order variation applies to all verbal MWLs, topicalisation of lexically fixed components is only rarely possible, as in G:*den Vogel dabei hat dann Jan abgeschossen* ('finally, Jan surpassed everyone', lit. the bird with it has then Jan shot).

3 IDAREX: Encoding Idioms As Regular EXpressions

The IDAREX formalism and the corresponding FSC Finite State Compiler have been developed at Rank Xerox Research Centre by L. Karttunen, P. Tapanainen and G. Valetto[2].

3.1 Morphological Variation

Because IDAREX uses a two-level morphology, words can be presented either in their base form at the lexical level or in an inflected form at the surface level. The *surface form* is preceded by a colon and restricts occurrences of the word to exactly this form, e.g.

```
:Welle
```

The *lexical form* is followed by an IDAREX morphological variable specifying morphological features of the word, and a colon, e.g.

```
durchschlagend A:
```

This represents any occurrence of the word with the specified morphological properties. The *morphological variable* can be very general, such as `'A'` for any adjectival use, or more specific, such as `Abse` for adjectives that may not be used in comparative form and `Nsg` to restrict nouns to the singular, as in

```
grün Abse:  Welle Nsg:
```

This way, the restricted morpho–syntactic flexibility of MWLs can be expressed very elegantly.

3.2 Modification

MWL modifications with particular words are represented as optional expressions with parentheses, as in

```
:den (:schönen) :Schein
```

The definition of *word–class variables* allows to express lexically unrestricted modifications of an MWL such as insertion of any adverb(s) (the *Kleene star operator* indicates that the item may occur any number of times):

```
perdre V: ADV* :la :tête
```

On the basis of simpler word–class variables more complex ones may be defined for complex syntactic categories such as NP, ADVP or PP.

3.3 Lexical and Structural Variation

The formalism provides a set of *RE operators* to combine the descriptions of single words. Square brackets '[ ]' and the bar '—' are used to describe lexical variants and alternations of more complex sequences such as word order variation in German. For instance, for the French example above we write

```
perdre V: ADV* [:la :tête | :la :boule
| :les :pédales ] ;
```

To express German verb–front and verb–final word order as in

'dabei wahrt er (immer) den Schein'
('in this, he always keeps up appearances')
and *'um den Schein zu wahren'*
('in order to keep up appearances')

we write

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

[2] For a more detailed description of the formalism see Karttunen and Yampol (1993), Tapanainen (1994), Karttunen (1995), and Segond and Tapanainen (1995).

In addition, IDAREX allows the definition of *macros* to capture generalisations on the syntactic level. Any position in the macro that we want to instantiate differently for each use is indicated by a parameter **$i**. Instantiations of parameters can be single words in lexical or surface form, variables, operators or other macros.

For example, instead of explicitly writing the complicated RE above, we define a word order macro **WOV1Arg** that may be used for all German verbal MWLs having no additional idiom–external arguments:

```
WOV1Arg:
  [ $2 Vfin:  (ADV* NPnom) ADV* $1
  | $1 (:zu) $2 V: ]
```

In addition, we define auxiliary macros **fix(i)** because we want to instantiate the parameter **$1**, which stands for the lexically fixed components of the MWL, with expressions of variable length:

```
fix5: $1 $2 $3 $4 $5    fix2: $1 $2    etc.
```

Using this word order macro, the MWLs *den (schönen) Schein wahren* and *die Ohren spitzen* ('to prick up one's ears', lit. the ears sharpen) can now both be expressed very simply according to the same schema as

```
WOV1Arg( fix5(:den ( :schönen )
         :Schein) wahren )
WOV1Arg( fix2(:die :Ohren) spitzen )
```

Further macros are defined for German for MWLs with a reflexive or particle verb, to express scrambling of an idiom–external PP complement or topicalisation. In French, macros describe for example the verb complex for MWLs involving a reflexive verb.

4 Discussion

NLP treatments of MWLs in so-called high level grammar formalisms have for example been proposed in Abeillé and Schabes (1989) in the framework of lexicalised TAGs, Erbach (1992) and Copestake (1994) in HPSG, Van der Linden (1993) in CG. These approaches to our knowledge cannot satisfactorily represent lexical variants, nor the restricted flexibility and modifiability of MWLs.

Instead of using a high–level grammar formalism we describe ,MWLs with finite–state local grammars. Although finite–state techniques are known to be unable to represent all the dependencies found in natural language, they have the advantage of allowing a very efficient treatment of a great number of phenomena and the implementation of robust, large–scale NLP systems. However, the use of these techniques is usually hampered by the unwieldiness in notation that these techniques usually lead to.

The presented approach overcomes this problem: instead of having to specify local grammars directly as finite state networks or as graphs (e.g. Maurel 1993, Roche 1993, and Silberztein 1993), IDAREX REs provide a convenient way to mix inflected and uninflected word forms, morphological features and complete word classes, thus greatly relieving lexicographers from the burden of explicitly listing all the possible forms. Furthermore, our formalism allows the use of a bigger set of operators such as **contains** (**$**), **not** (~), **and** (&), etc. This provides us with the possibility to express certain things in a very compact way. For example, in the definition of German verbs we exclude contracted forms of verbs and the pronoun *es* such as *geht's* ('it goes') , we simply state "any expression with the morphological feature **+V**, followed by anything that must not contain a letter (i.e. additional morphological features), and which does not contain the feature **+es** in any position"

```
define V   %+V ~$Letter & ~$%+es
```

With macros, generalizations about patterns that can occur for a whole class of MWLs can be expressed. This compactness and flexibility are, as far as we know, specific to our approach.

Encoding the local grammars as REs instead of encoding them directly as networks does of course not change the expressive power of the formalism, but it conveniently abstracts the handling of MWLs from the graph manipulation level, allowing to develop and employ devices that operate on string representations and map them to the underlying finite state networks. As we have shown above, this simplifies considerably the description of the different patterns of variation occuring in MWLs.

Once the MWLs listed in the dictionary have been manually changed into their ***canonical base form***, including possible lexical variants and modifiers and indicating morphologically flexible components and the scope of alternative components, the IDAREX REs describing all possible contexts in which the MWLs can occur can be produced automatically. For instance, the canonical forms for the examples from section 2 can be specified as:

durchschlagender° Erfolg°
grüne° Welle (sg)
perdre° (ADV)^la tête/la boule/ les pédales^
T: den Vogel (bei etw) abschießen°
den (schönen) Schein wahren°

das Handtuch werfen°
out of the blue
um Haaresbreite

Such canonical base forms, somewhat similar in spirit to the notation used in Longman's 'Dictionary of English Idioms' (1979), do not only form the basis for the automatic processing and recognition of MWLs. Human users as well would profit from a careful description of the variability of MWLs, so it should be worthwhile to also include the canonical forms in dictionaries for human users.

The presented approach is successfully used in the COMPASS project[3] to represent MWLs in dictionary databases converted from standard bilingual dictionaries. The COMPASS system, based on the LOCOLEX engine (Bauer, Segond and Zaenen 1995) developed at RXRC, allows look–up of words in the dictionary database directly out of an on–line text. When the user clicks on an unknown word in a foreign language, LOCOLEX evaluates the context of the queried word. Currently, the system determines the word's part of speech and whether the word is part of an MWL. In the latter case, the translation for the entire MWL is returned, otherwise a selection of translations for the most appropiate part of speech.

Acknowledgments

We thank Annie Zaenen, Lauri Karttunen, Ted Briscoe, and Irene Maxwell for their comments on an earlier draft of this paper.

References

Abeillé Anne ; and Schabes Yves, (1989). "Parsing idioms in lexicalized TAGs". *Proceedings of the 4th EACL*, Manchester, UK.

Bauer Daniel ; Segond Frédérique ; and Zaenen Annie, (1995). "LOCOLEX: the translation rolls off your tongue". *Proceedings of ACH-ALLC*, Santa Barbara, CA.

Brundage Jennifer ; Kresse Maren ; Schwall Ulrike ; and Storrer Angelika, (1992). "Multiword Lexemes: A Monolingual and Contrastive Typology for NLP and MT". *IWBS Report 232*, IBM TR-80.92-029, IBM Deutschland GmbH, Institut für Wissensbasierte Systeme, Heidelberg, September.

Copestake Anne, (1994). "Idioms in general and in HPSG". Presentation given at the Workshop *'The Future of the Dictionary'*, Uriage-Les-Bains, France, September 1994.

Engelke Sabine, (1994). *Eigenschaften von Phraseolexemen: Eine Untersuchung zur syntaktischen Variabilität und internen Modifizierbarkeit von somatischen verbalen Phraseolexemen.* Master's Thesis, Universität Tübingen, Germany, April.

Erbach Gregor, (1992). "Head-Driven Lexical Representation of Idioms in HPSG". In M. Everaert et al., editors, *Proceedings of the International Conference on Idioms*, Tilburg, NL, September.

Fleischer Wolfgang, (1982). *Phraseologie der deutschen Gegenwartssprache.* VEB Bibliographisches Institut, Leipzig, Germany.

Gross Maurice. (1982). "Une classification de phrases figées français". *Revue Quebecoise de Linguistique*, Vol. 11, No. 2. Montreal.

Karttunen Lauri, (1995). "The Replace Operator". *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Boston, MA.

Karttunen Lauri ; and Yampol Todd, (1993). "Interactive Finite-State Calculus". *Technical Report ISTL-NLTT-1993-04-01*, Xerox Palo Alto Research Center, California.

van der Linden Erik-Jan, (1993). *A Categorial Computational Theory of Idioms.* OTS Dissertation Series, Utrecht, NL.

Maurel Denis, (1993). "Passage d'un automate avec tables d'acceptablilité à un automate lexical". In *Actes du colloque Informatique et langue naturelle*, pages 269–279, Nantes, France.

Nunberg Geoffrey ; Wasow Thomas ; and Sag Ivan, (1994). "Idioms". *Language*, 70/3:491–538.

Roche Emmanuel, (1993). *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire.* Thèse de doctorat, Université Paris 7.

Segond Frédérique ; and Breidt Elisabeth, (1995). "Description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis". *Lexicomatique et Dictionnairiques*, Actes des IVe Journées Scientifiques du reseau "Lexicologie, Terminologie, Traduction" de l'UREF, Lyon, Septembre 1995.

Segond Frédérique ; and Tapanainen Pasi, (1995). "Using a finite-state based formalism to identify and generate multiword expressions". *Technical Report MLTT-019*, Rank Xerox Research Centre, Grenoble, France, July.

[3] 'Adapting bilingual dictionaries for COMPrehension ASSistance', LRE-62-080.

Silberztein Max, (1993). *Dictionnaires électroniques et analyse automatique de textes – Le système INTEX.* Masson, Paris, France.

Tapanainen Pasi, (1994). "RXRC Finite-State rule Compiler". *Technical Report MLTT-020*, Rank Xerox Research Centre, Grenoble, France.