# Multiword Units in an MT Lexicon

**Tamás Váradi**
Linguistics Institute
Hungarian Academy of Sciences
varadi@nytud.hu

## Abstract

Multiword units significantly contribute to the robustness of MT systems as they reduce the inevitable ambiguity inherent in word to word matching. The paper focuses on a relatively little studied kind of MW units which are partially fixed and partially productive. In fact, MW units will be shown to form a continuum between completely frozen expression where the lexical elements are specified at the level of particular word forms and those which are produced by syntactic rules defined in terms of general part of speech categories. The paper will argue for the use of local grammars proposed by Maurice Gross to capture the productive regularity of MW units and will illustrate a uniform implementation of them in the NooJ grammar development framework.

## 1 Introduction

The robustness of MT systems crucially depend on the size and quality of their lexical componenets. It is commonly recognized that word-to-word equivalents are fraught with ambiguities. MW units on the other hand carry, as it were, the disambiguating context with them. Hence, the more MW units in the lexicon and the longer they are, the less noisy and more robust the MT lexicon is likely to be. However, not all kinds of MW units are amenable to inclusion by itemized listing in the lexicon. The paper will focus on MW units whose structure contains slots that can be filled by more or less open ended lexical units. They are treated in paper dictionaries with the usual method of exemplification and implication, which, even if the intended extension of the set of expression is clear, is obviously not a viable option in a machine system that cannot rely on the linguistic competence and world knowledge that human readers of dictionaries are expected to bring to the job of interpreting lexical entries.

## 2 The multi-word unit continuum

In order to develop first an intuitive grasp of the phenomena, consider the following examples.

```
1)  English-speaking population
    French-speaking clients
    Spanish-speaking students
```

It would not be difficult to carry on with further examples, each embodying a pattern <language-name> speaking <person> or <group of persons>. It is a prototypical example for our purposes because the words are interdependent yet they admit of open-choice in the selection of lexical items for certain positions. The phrases *speaking students, English-speaking, or English population are either not well-formed or does not mean the same as the full expression. The meaning of the phrase is predominantly, if perhaps not wholly, compositional and for native language speakers the structure may seem entirely transparent. However, in a bilingual context this transparency does not necessarily carry over to the other language. For example, the phrases in (1) are expressed in Hungarian as in 2)

```
2)Angol nyelvű          lakosság
  English language-Adj  population

  Fracia nyelvű         ügyfelek
  French language-Adj   clients

  Spanyol nyelvű        diákok
   Spanish language-Adj  students
```

The Hungarian equivalent bears the same characteristics of semantic compositionality and structural transparency and is open-ended in the same points as the corresponding slots in the English pattern. It would be extremely wasteful to capture the bilingual correspondences in an itemized manner, particularly as the set of expressions on both sides are open-ended anyway.

At the other end of the scale in terms of productivity and compositionality one finds phrases like those listed in 3)

```
3) English breakfast
   French fries
   German measles
```

Purely from a formal point of view, the phrases in 3) could be captured in the pattern `<language name><noun>` but the co-occurrence relations between items in the two sets are limited to the extreme so that once they are defined properly, we are practically thrown back to the particular one-to-one combinations listed in 3).

Note that if we had a set like 4), where one element is shared it would still not make sense make sense to factorize the shared word *French* because it enters into idiomatic semantic relations. In other words, the multi-word expressions are semantically non-compositional even in terms of English alone.

```
4) French bread
   French horn
   French dressing
```

The set of terms in 5) exemplifies the other end of the scale in terms of compositionality and syntactic transparency. They are adduced here to exemplify fully regular combinations of words in their literal meaning.

```
5) French schools
   French vote
   French books
   French drivers
```

In between the wholly idiosyncratic expressions which need to be listed in the lexicon and the set of completely open-choice expressions which form the province of syntax, there is a whole gamut of expressions that seem to straddle the lexicon-syntax divide. They are non-compositional in meaning to some extent and they also include elements that come from a more or less open set. Some of these open-choice slots in the expressions may be filled with items from sets that are either infinite (like numbers) or numerous enough to render them hopeless or wasteful for listing in a dictionary. For this reason, they are typically not fully specified in dictionaries, which have no of means of representing them explicitely in any other way than by listing. For want of anything better, lexicographers rely on the linguistic intelligence of their readers to infer from a partial list the correct set of items that a given lexical unit applies to. Bolinger (Bolinger 1965) elegantly sums up this approach as

> Dictionaries do not exist to define, but to help people grasp meaning, and for this purpose their main task is to supply a series of hints and associations that will relate the unknown to something known.

Adroit use of this technique may be quite successful with human readers but is obviously not viable for NLP purposes. What is needed is some algorithmic module in order to model the encoding/decoding processing that humans do in applying their mental lexicon. The most economical and sometimes the only viable means to achieve this goal is to integrate some kind of rule-based mechanism that would support the recognition as well as generation of all the lexical units that conventional dictionaries evoke through well-chosen partial set of data.

## 3 Local grammars

Local Grammars, developed by Maurice Gross (Gross 1997), are heavily lexicalized finite state grammars devised to capture the intricacies of local syntactic or semantic phenomena. In the mid-nineties a very efficient tool, INTEX was developed at LADL, Paris VII, (Silberztein 1999) which has two components that are of primary importance to us: it contains a complex lexical component (Silberztein 1993) and a graphical interface which supports the development of finite state transducers in the form of graphs (Silberztein 1999).

Local grammars are typically defined in graphs which are compiled into efficient finite state automata or transducers. Both the lexicon and the grammar are implemented in finite state transducers. This fact gives us the ideal tool to implement the very kind of lexicon we have been arguing for, one that includes both static entries and lexical grammars.

The set of expressions discussed in 1) can be captured with the graph in Figure 1. It shows a simple finite state automaton of a single with through three nodes along the way from the initial symbol on the left to the end symbol on the right. It represents all the expressions that match as the graph is traversed between the two points. Words in angle brackets stand for the lemma form, the shaded box represent a subgraph that can freely be embedded in graphs. The facility of

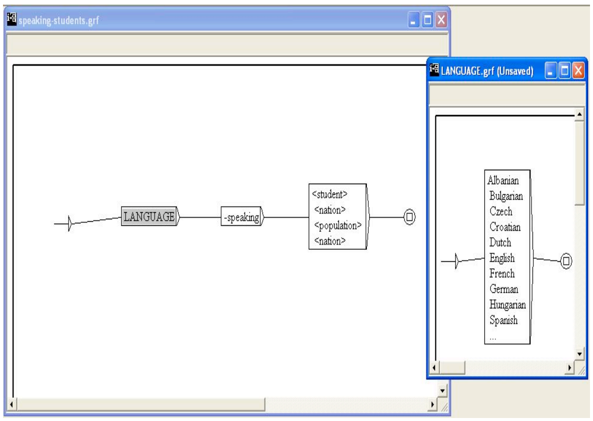

**Figure 1** INTEX/NOOJ graph to capture phrases like *English-speaking students*

graph embedding has the practical convenience that it allows the reuse of the subgraph in other contexts. At a more theoretical level, it introduces the power of recursion into grammars. Subgraphs may also be used to represent a semantic class, such as language name in the present case, and can be encoded in the dictionary with a semantic feature like +LANGNAME. INTEX/NOOJ dictionaries allow an arbitrary number of semantic features to be represented in the lexical entries and they can be used in the definition of local grammars as well. An alternative grammar using semantic features is displayed in Figure 2.

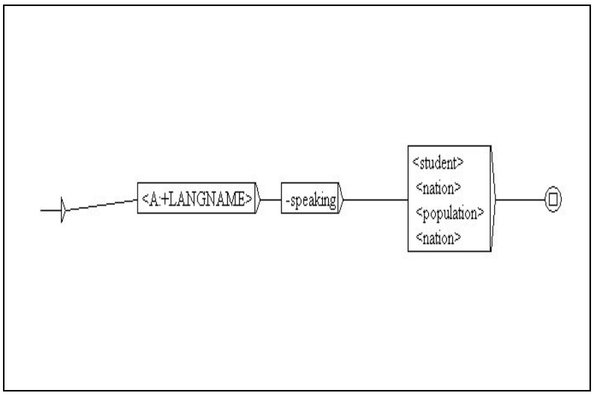

**Figure 2** Representing the phrases in Figure 1 with semantic features

Note that to render expressions like in 2) we use local grammars containing nodes that range from specific word forms through lemmas, lists of words, words defined by a semantic class in an ontology to syntactic class or even the completely general placeholder for any word. Such flexibility allows us to apply the constraint defined at the right level of generality required to cover exactly the set of expressions without overgeneration.

The local grammars defining the kind of partially productive multi-word units that the present paper focuses on can typically be defined with the nodes being defined in terms of some natural semantic class such as the language names of examples 2) or names of colours or body parts illustrated in 6)

```
6a) the lady in black
6b) a fekete ruhás hölgy
    the black clad lady
```

The English expression in 6a) can be implemented with the graph in Figure 3, its Hungarian equivalent 6b) is displayed in Figure 4.

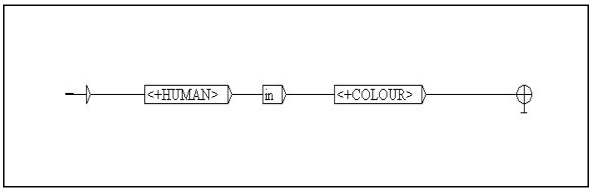

**Figure 3** Local grammar to cover the expressions like 6a)

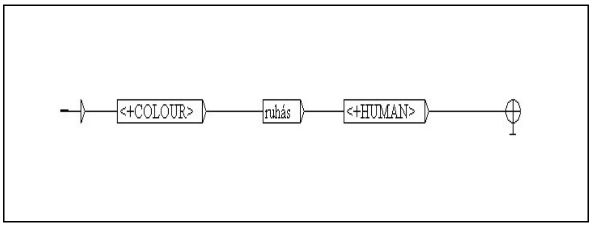

Figure 4 Local grammar to cover the expressions like 6b)

The use of semantic features is merely the first step in building an efficient lexicon. At a more advanced level, the lexicon would include a system of semantic features arranged into typed hierarchy, which would allow use of multiple inheritence.

## 4 Application of local grammars

In the present section we provide some examples of how rendering multi-word units with local grammars can enhance a multi-lingual application.

### 4.1 Semantic disambiguation

The use of transducers in INTEX/NOOJ provides an intuitive and user-friendly means of semantic disambiguation as illustrated in Figure 5. Here the appropriate meaning of the specific node is defined by its Hungarian equivalent, but of course one might just as well have used monolingual tags for the same purpose.

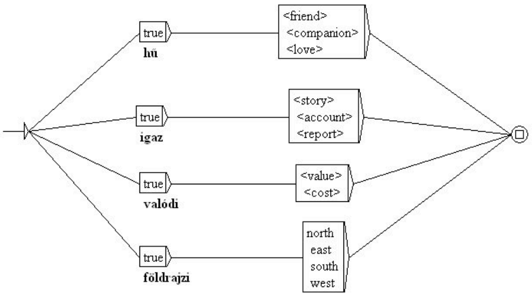

**Figure 5**. Semantic disambiguation with the use of local grammars

### 4.2 Partial automatic translation

On the analogy of shallow parsing, we may compile transducers that produce as output the target language equivalent of the chunks recognized. This is illustrated in Figure 6 where the expressions “trade/trading in dollar/yen” etc. are rendered as “dollárkereskedelem, jenkereskedelem” etc. whereas “trade/trading in Tokyo/London” etc. are translated as “tókiói/londoni kereskedés”. Note that the recognized words are stored in a variable captured by the labelled brackets and used in the compilation of the output.

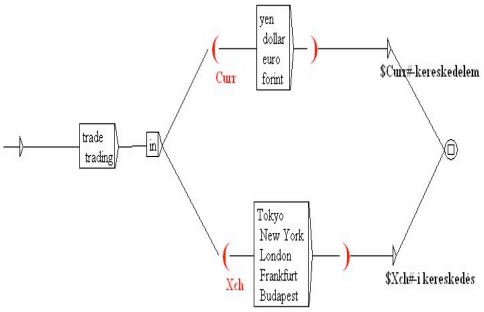

**Figure 5** Partial translation transducers using variables

### 4.3 Automatic lexical acquisition

Local grammars can be used not only for recognition and generation but also for automated lexical acquisition. This can be achieved by suitably relaxing the constraints on one or more of the nodes in a graph and apply it to a large corpus. The resulting hit expressions can then be manually processed to find the semantic feature underlying the expressions or establish further subclasses etc.

As an example, consider Figure 7 containing a graph designed to capture expressions describing various kinds of *graters* in English. As Figure 6 shows the entry for *grater* in the Oxford Advanced dictionary (Wehmeier 2005) uses only hints through specific examples as to what sort of graters there may be in English

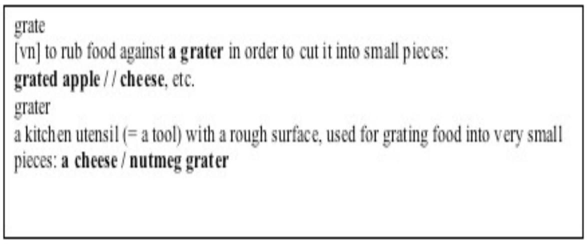
grate
[vn] to rub food against **a grater** in order to cut it into small pieces:
**grated apple / / cheese**, etc.
grater
a kitchen utensil (= a tool) with a rough surface, used for grating food into very small pieces: **a cheese / nutmeg grater**

**Figure 6** Part of the dictionary entry *GRATE* from OALD7

The node <MOT> matches an arbitrary word in INTEX, the symbol <E> covers an empty element, used here in disjunction the syntactic category <DET> to turn the latter optional.

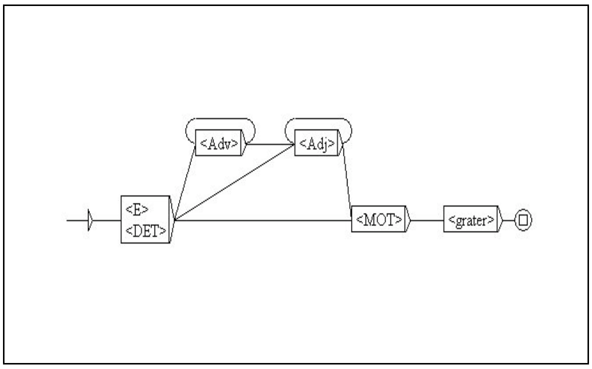

**Figure 7** Automatic aquisition of multi-word units with local grammars

## 5 Conclusions

In the present paper we have highlighted the importance of multi-word units that are partially productive. Far from being peripheral, they appear to be ubiquitous particularly when viewed in a multilingual setting. Many of these expressions including such common phrases like `a twenty year old woman` may not be viewed as multi-word expressions at all until one realizes the syntactic/semantic constraints involved in their structure (e.g. `*year old woman`). More importantly, once their translation to another language is not entirely transparent (i.e. they cannot be rendered word by word), the crosslingual transfer must be registered. It is suitably done in traditional dictionaries through a single example, but in an MT system such reliance on the active contribution of the human user is not an option. Nor is exhaustive listing, as proved by this simple but extremely common example.

We have shown how the use of local grammars can provide the flexibility required to cover the phenomena of partially productive multiword units which form a continuum between frozen multi-word expressions and open-ended productive phrases defined by syntactic rules sensitive to part of speech categories only.

The local grammars were illustrated in some multilingual applications using the grammar development environment INTEX/NOOJ, which provide an intuitive and linguistically sophisticated tool to explore the use of the multi-word units in question.

## References

Bolinger, D. (1965). "The Atomization of Meaning." Language **41**: 555-573.

Gross, M. (1997). The Construction of Local Grammars. in Y. S. Emmanuel Roche (szerk.) Finite State Language Processing. MIT Press**:** 329-352.

Sag, I. et al. 2002 Multiword Expressions: A Pain in the Neck for NLP. in Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics CICLING 2002): 1--15,

Silberztein, M. (1993). Dictionnaires électorniques et analyse automatique de textes: le systeme INTEX. Paris, Masson.

Silberztein, M. (1999). "Text Indexation with INTEX." Computers and the Humanities **33**(3): 265-280.

Wehmeier, S., (szerk.) (2005). Oxford Advanced Learner's Dictionary. Oxford, Oxford University Press.