Accommodating Multiword Expressions in an Arabic LFG Grammar

Mohammed A. Attia

School of Informatics, The University of Manchester, UK
mohammed.attia@postgrad.manchester.ac.uk

Abstract. Multiword expressions (MWEs) vary in syntactic category, structure, the degree of semantic opaqueness, the ability of one or more constituents to undergo inflection and processes such as passivization, and the possibility of having intervening elements. Therefore, there is no straight-forward way of dealing with them. This paper shows how MWEs can be dealt with at different levels of analysis starting with tokenization, and going through the stages of morphological analysis and syntactic parsing.

1 Introduction

There was a tendency to ignore MWEs in linguistic analysis due to their complexity and idiosyncrasy. However, it is now recognized that MWEs have a high frequency in day-to-day interactions (Venkatapathy, 2004), that they account for 41% of the entries in WordNet 1.7 (Fellbaum, 1998, Sag et al., 2001), that phrasal verbs account for "about one third of the English verb vocabulary" (Li et al., 2003), and that technical domains rely heavily on them. This makes it imperative to handle MWEs if we want to make large-scale, linguistically-motivated, and precise processing of the language.

MWEs constitute serious pitfalls for machine translation systems and human translators as well (Volk, 1998). When they are translated compositionally, they give textbook examples of highly intolerable, blind and literal translation. It is also underestimation to the problem to assume that it should be handled during higher phases of processing such as transfer. In fact MWEs require deep analysis that starts as early as the tokenization, and goes through morphological analysis and into the syntactic rules. The focus of this paper is to explain how MWEs can be accommodated in each step in the preprocessing and the processing stages. The advantages of handling MWEs in the pre-processing stage are avoidance of needless analysis of idiosyncratic structures, reduction of parsing ambiguity, and reduction of parse time (Brun, 1998). This is why there are growing calls to construct MWE dictionaries (Guenthner and Blanco, 2004), lexicons (Calzolari et al., 2002), and phrasets (Bentivogli and Pianta, 2003).

This paper shows how several devices can be applied to handle MWEs properly at several stages of processing. All the solutions are applied to Arabic, yet, most of the solutions are general and are applicable to other languages as well. The software used for writing grammar rules is XLE (Xerox Linguistic Environment) (Butt et al., 1999,

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 87 – 98, 2006.

Dipper, 2003). It is a platform created by Palo Alto Research Center (PARC) for developing large-scale grammars using LFG (Lexical Functional Grammar) notations. Morphological transducers, tokenizers and MWE transducers are all developed using Finite State Technology (Beesley and Karttunen, 2003).

2 Definition

MWEs encompass a wide range of linguistically related phenomena that share the criterion of being composed of two words or more, whether adjacent or separate. MWEs have been defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2001). In an MWE, the structure and the semantics of the expression are dependant on the phrase as a whole, and not on its individual components (Venkatapathy, 2004). MWEs cover expressions that are traditionally classified as idioms (e.g. *down the drain*), prepositional verbs (e.g. *rely on*), verbs with particles (e.g. *give up*), compound nouns (e.g. *book cover*) and collocations (e.g. *do a favour*).

Although there is no clear-cut definition with which we can decide what expressions can be considered MWEs, there is a set of criteria (adapted from (Baldwin, 2004, Calzolari et al., 2002, Guenthner and Blanco, 2004)) when one or more of which applies the expression can safely be considered as an MWE.

1. Lexogrammatical fixedness. The expression has come to a rigid or frozen state. This fixedness can be identified through a number of tests. Components of the expression must be immune to the following operations:
   - Substitutability. The word *many* in (1.a) can be substituted with its synonym *several*, while in (1.b) it cannot.
     (1.a) *many books -> several books* (1.b) *many thanks -> * several thanks*
   - Deletion. The adjective in (2.a) can be deleted, while in (2.b) it cannot.
     (2.a) *black sheep -> the sheep* (2.b) *black hole -> * the hole*
   - Category transformation. The adjective in (3.a) can be changed to noun, while in (3.b) it cannot.
     (3.a) *black sheep -> the blackness of the sheep*
     (3.b) *black hole -> * the blackness of the hole*
   - Permutation. A noun-noun compound can usually be expressed by a noun-preposition-noun as in (4.a), but not in MWEs as in (4.b) and (4.c).
     (4.a) *the hospital staff -> the staff of the hospital*
     (4.b) *life guard -> * the guard of life* (4.c) *kiss of life -> * life kiss*
2. Semantic non-compositionality. The meaning of the expression is not derived from the component parts, such as *kick the bucket* which means *die*.
3. Syntactic irregularity. The expression exhibits a structure that is inexplicable by regular grammatical rules, such as *long time, no see* and *by and large*.
4. Single-word paraphrasability. The expression can be paraphrased by a single word, such as *give up* which means *abandon*.
5. Translatability into a single word or into a non-compositional expression. Expressions can be considered as "terms when the corresponding … translation is a unit, or when their translation differs from a word to word translation" (Brun, 1998). In various projects a corpus of translated texts is used to judge or detect

MWEs (Butt et al., 1999, Nerima et al., 2003, Smadja et al., 1996). Sometimes a unilingual analysis may be confused about whether an expression is a regular combination of words or an MWE. Translation usually helps to show expressions in perspective.

(5) *looking glass* = مرآة *[mir'aah]* (Arabic)

3 Classification of Multiword Expressions

In order for an expression to be classified as an MWE, it must show a degree of semantic non-compositionality and/or a degree of morpho-syntactic inflexibility. MWEs are classified with regard to their semantic compositionality into lexicalized and institu-tionalized expressions. Moreover, they are classified with regard to their flexibility into fixed, semi-fixed and syntactically flexible expressions (adapted from (Sag et al., 2001)).

3.1 Compositional vs. Non-compositional MWEs

Semantic compositionality, sometimes termed *decomposability*, is "a means of describing how the overall sense of a given idiom is related to its parts" (Sag et al., 2001). An illustrative example of non-compositionality is the expression *kick the bucket*, where the meaning "die" has no relation to any word in the expression. An example of compositional expressions is the compound noun *book cover*, where the meaning is directly related to the component parts. Unfortunately, the assignment of a plus/minus feature of compositionality to an expression is sometimes very elusive. Most of the time "one cannot really make a binary distinction between compositional and non-compositional MWEs" (Venkatapathy, 2004). They occupy a continuum in a large scale. At one end of the scale there are those expressions that are highly opaque and non-compositional, where the meaning is not traceable to any of the component parts, such as *kick the bucket*. In the middle of the scale there are those where one or more words are used in an idiosyncratic sense, or use "semantics unavailable outside the MWE" (Baldwin et al., 2003), such as *spill the beans*, which means "to disclose a secret". At the other end of the scale there are those which are highly compositional, such as *book cover*, *traffic light*, *health crisis* and *party meeting*.

Non-compositional expressions, or, more accurately, expressions that show any degree of non-compositionality, are termed *lexicalized* and are automatically eligible to be considered as MWEs. However, in order for compositional expressions to be included in an MWE lexicon, they need to be conventionalized or *institutionalized*. This means that these expressions have come to such a frequent use that they block the use of other synonyms and near synonyms (Nerima et al., 2003). When words co-occur in a statistically meaningful way like this they are called *collocations*. This way, expressions such as *book cover* and *traffic light* can be safely added to an MWE lexicon, while *health crisis* and *party meeting* cannot.

3.2 Flexible vs. Inflexible MWEs

With regard to syntactic and morphological flexibility, MWEs are classified into three types: fixed, semi-fixed and syntactically flexible expressions (Baldwin, 2004, Oflazer et al., 2004, Sag et al., 2001).

3.2.1 Fixed Expressions

These expressions are lexically, syntactically and morphologically rigid. An expression of this type is considered as a word with spaces (a single word that happens to contain spaces), such as *San Francisco* and *in a nutshell*. Some expressions are frozen at the level of the sentence, sometimes termed "frozen texts" (Guenthner and Blanco, 2004). These include proverbs such as *Buy cheap, buy twice*, and pragmatically fixed expressions such as *Good morning*.

3.2.2 Semi-fixed Expressions

These expressions can undergo morphological and lexical variations, but still the components of the expression are adjacent. They cannot be reordered or separated by external elements. The variations that can affect semi-fixed expressions are of two types:

1. Morphological variations that express person, number, tense, gender, etc., such as *traffic light/lights* and *kick/kicks/kicked the bucket*.
2. Lexical variations. This is the case when a position in the expression is filled by a choice from the set of reflexive pronouns (e.g. *prostrate himself/herself*), or when one word can be replaced by another (e.g. *to sweep something under the carpet/rug*).

3.2.3 Syntactically Flexible Expressions

These are the expressions that can either undergo reordering, such as passivization (e.g. *the cat was let out of the bag*), or allow external elements to intervene between the components (e.g. *slow the car down*). Here the adjacency of the MWE is disrupted.

4 Handling MWEs

This section shows how an MWE transducer is built to complement the morphological transducer, and how the MWE transducer interacts with other processing and preprocessing components. It also shows how the grammar is responsible for detecting and interpreting syntactically flexible expressions.

4.1 Building the MWE Transducer

A specialized two-sided transducer is build for MWEs using a finite state regular expression (Beesley and Karttunen, 2003) to provide correct analysis on the lexical side (upper side) and correct generation on the surface side (lower side). This transducer covers two types of MWEs: fixed and semi-fixed expressions, leaving syntactically-flexible expressions to be handled by the grammar. This entails that the

MWE transducer will not handle verbs at all (in the case of Arabic), and will not handle compound nouns that allow external elements to intervene. In order for the transducer to account for the morphological flexibility of some components, it consults the core morphological transducer (Attia, 2005) to obtain all available forms of certain words. This is how the MWE is enabled to search through the core morphological transducer. First the morphological transducer is loaded and put in a defined variable:

(6) load ArabicTransducer.fst
define AllWords

For the word وزير (wazir [minister]), for instance, the transducer has the following upper and lower structures.

(7) +noun [وزير]+masc+sg
وزير

In order to capture all different forms of the word (number and gender variations) we compose the rule in (11) above the finite state network (or transducer).

(8) $[?* "[" {وزير} "]" ?*] .o. AllWords

The sign "$", in finite state notations, means only paths that contain the specified string, and "?*" is a regular expression that means any string of any length. This gives us all surface forms that contain the wanted stem.

Arabic Multiword Nouns

Fixed compound nouns are entered in the lexicon as a list of words with spaces. Example (9) shows how the compound noun حفظ الأمن (hifz al-amn [peace keeping]) is coded in a finite state regular expression.

(9) ["+noun" "+masc" "+def"]:{حفظ} sp {الأمن}

The string "sp" here indicates a separator or space between the two words, so that each word can be identified in case there is need to access it. Compound proper names, including names of persons, places and organizations, are treated in the same way.

Semi-fixed compound nouns that undergo limited morphological/lexical variations are also entered in the lexicon with the variations explicitly stated. Example (10) shows the expression نزع سلاح (naz' silah [lit. removing a weapon: disarming]) which can have a definite variant.

(10) ["+noun" "+masc"]:{نزع} sp ("+def":{ال}) {سلاح}

Example (11) illustrates lexical variation. The expression مدعى عليه (mudda'a 'alaih [lit. the charged against him: defendant]) can choose from a fixed set of third person pronouns to indicate the number and gender of the noun.

(11)["+noun"]:0 ("+def":{ال}) {مدعى} sp {علي} ["+sg" "+masc":ه
|"+sg" "+fem":{ها}|"+dual":{هما} | "+pl" "+masc":{هم} | "+pl" "+fem":{هن}]

As for Semi-fixed compound nouns that undergo full morphological variations, a morphological transducer is consulted to obtain all possible variations.

First we need to explain how Arabic compound nouns are formed and what morphological variations they may have. They are generally formed according to the re-write rule in (12).

(12) NP[_Compound] -> [N N* A*] & ~N

This means that a compound noun can be formed by a noun optionally followed by one or more nouns, optionally followed by one or more adjectives. The condition "&~N" is to disallow the possibility of a compound noun being composed of a single noun. In an N_N construction, the first noun is inflected for number and gender, while the second is inflected for definiteness. When the compound noun is indefinite there is no article attached anywhere, but when it is definite, the definite article ال (al [the]) is attached only to the last noun in the structure. The compound وزير الخارجية (wazir al-kharijiyah [foreign minister]) is formatted as in (13).

(13) $[?* "[" {وزير} "]" ?*] .o. AllWords sp ("+def":{ال}) {خارجية}

In an N_A structure the noun and adjective are both inflected for number and gender and can take the definite article. The regular expression in (14) shows the format of the expression سيارة مفخخة (saiyarah mufakhakhah [lit. trapping car: car bomb]).

(14) $[?* "[" {سيارة} "]" ?*] .o. AllWords sp $[?* "[" {مفخخ} "]" ?*] .o. AllWords

This regular expression, however, is prone to overgenerate allowing for a masculine adjective to modify a feminine noun in contradiction to agreement rules. This is why paths need to be filtered by a set of combinatorial rules (or local grammars). The rules in (15) discard from the network paths that contain conflicting features:

(15) ~$["+dual" <> ["+sg" | "+pl"] /?*] .o. ~$["+fem" <> "+masc" /?*]

The notation "~$" means "does not contain," "<>" means "order is not important" and "/?*" means "ignore noise from any intervening strings".

After the words are combined correctly, they need to be analyzed correctly. First we do not need features to be repeated in the upper language. In example (16.a), the noun سيارة (saiyarah [car]) is analayzed as +fem+sg, and the adjective مفخخة (mufakhakhah [trapping]) has the same features +fem+sg. Second we do not want features to be contradictory. The first word is analyzed as +noun, while the second is analyzed as +adj. This is shown by the representation in (16.b).

(16.a) سيارة مفخخة

saiyarah	mufakhakhah
car.noun.fem.sg	trapping.adj.fem.sg (bomb car)

(16.b)

+noun+fem+sgسيارة	+adj+fem+sg مفخخة
سيارة	مفخخة

We need to remove all features from non-head components, and the rules in (17) serve this purpose.

(17) "+sg" -> [] || sp ?* _ .o. "+fem" -> [] || sp ?* _
.o. "+adj" -> [] || sp ?* _ .o. "+noun" -> [] || sp ?* _

When these rules are applied to the upper language in the transducer, they remove all specified features from non-initial words, leaving features unique and consistent.

(18)

+noun+fem+sgسيارة	مفخخة
سيارة	مفخخة

Special attention, however, should be given to cases where some features are drawn from non-initial nouns like definiteness in (13) above and the features of number and gender in (11).

Adjectives, Adverbs and Others

Adjectives are treated to a great extent like semi-fixed expressions, as they can undergo morphological variations, such as the examples in (19).

(19.a)	قصير النظر	(19.b)	قصيرات النظر
	qasir al-nazar		qasirat al-nazar
	short.masc.sg sighted		short.fem.pl sighted

Some adverbs have regular forms and can be easily classified and detected. They are usually composed of a preposition, noun and a modifying adjective. The preposition and the noun are relatively fixed while the adjective changes to convey the meaning, as shown by (20).

(20) بطريقة عشوائية (bi-tariqah ‘ashwa’iyah [randomly / lit.: in a random way])

Some MWEs, however, are less easily classified. They include expressions that function as linking words, as in (21), and highly repetitive complete phrases as in (22).

(21) وعلى هذا (wa-‘ala haza [whereupon])

(22) ومما يذكر أن (wa mimma yuzkar anna [It is to be mentioned that])

One String MWEs

Some MWEs in Arabic are composed of words with clitics. They look like single words but if they are to be treated by the morphological analyzer alone, they will be analyzed compositionally and lose their actual meaning and syntactic function, such as the example in (23).

(23) بالتالي (bit-tali [consequently / lit.: with the second])

4.2 Interaction with the Tokenizer

The function of a tokenizer is to split a running text into tokens, so that they can be fed into a morphological transducer for processing. The tokenizer is responsible for demarcating words, clitics, abbreviated forms, acronyms, and punctuation marks. The output of the tokenizer is a text with a mark after each token; the “@” sign in XLE case. Besides, the tokenizer is responsible for treating MWEs in a special way. They should be treated as single tokens with the inner space(s) preserved.

One way to allow the tokenizer to handle MWEs is to embed them in the Tokenizer (Beesley and Karttunen, 2003). Yet a better approach, described by (Karttunen et al., 1996), is to develop one or several multiword transducers or “staplers” that are composed with the tokenizer. I will explain here how this is implemented in my solution. Let’s first look at the composition regular expression:

```
(24)  1   singleTokens.i
      2   .o. ?* 0:"[[[" (MweTokens.l) 0:"]]]" ?*
      3   .o. "@" -> " " || "[[[" [Alphabet* | "@"*] _
      4   .o. "[[[" -> [] .o. "]]]" -> [].i;
```

The tokenizer is defined in the variable *singleTokens* and the MWE transducer is defined in *MweTokens*. In the MWE transducer all spaces in the lower language are replaced by “@” so that the lower language can be matched by the output of the tokenizer. In line 1 the tokenizer is inverted (the upper language is shifted down) by the operator “.i” so that composition goes in the right direction. From the MWE

transducer we take only the lower language by the operator ".l" in line 2. Here all MWEs are searched and if they match any string they will be enclosed with three brackets on either side. Line 3 replaces all "@" signs with spaces in MWEs only. The two compositions in line 4 remove the intermediary brackets.

Let's now show how this works with an example:

(25) ولوزير خارجيتها
wa-li-wazir kharijiyati-ha
and-to-minister foreign-its (and to its foreign minister)

The tokenizer first gives the output in (26), among other possibilities:

(26) @و@ل@وزير@خارجية@ها@ (approx. and@to@foreign@minister@its@)

Then after the MWEs are composed with the tokenizer, we obtain the result in (27) with the MWE identified as a single token:

(27) @و@ل@وزير خارجية@ها (approx. and@to@foreign minister@its@)

4.3 Integration with the Morphological Transducer

The MWE transducer can either complement or substitute the core morphological transducer. If we want to allow the compositional analysis of the expression to be available to the parser we need make the MWE transducer complement the morphological transducer. On the other hand if we are sure enough that MWEs cannot have significant compositional varieties, we need to prioritize the MWE transducer over the main transducer, so that when an expression is found in the MWE transducer no further search is done.

4.4 Interaction with the Grammar

As for fixed and semi-fixed MWEs that are identified both by the tokenizer and the morphological analyzer, they are represented in Lexical Functional Grammar (LFG) as a single word, as shown in (28).

(28.a) جنود حفظ الأمن (junud hifz al-amn [peace keeping soldiers])

(28.b) C-Structure

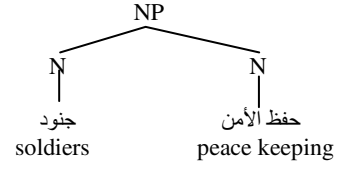

Fig. 1. C-structure of an MWE NP

(28.c) F-Structure

SUBJ PRED 'جنود[soldiers]'
MOD [PRED 'حفظ الأمن[peace keeping]'
DEF +, GEND masc, NUM sg, PERS 3]
DEF +, GEND masc, NUM pl, PERS 3

Fig. 2. F-structure of an MWE NP

When MWEs are syntactically flexible, by either allowing reordering such as passivization or allowing intervening elements such as phrasal verbs, they are handled by the syntactic parser. As passivization in Arabic is not made by configurational restructuring of the sentence, but rather by morphological inflection of verbs, we can say that Arabic shows only one instance of syntactic flexibility in MWEs, that is allowing intervening elements.

Syntactically flexible MWEs are handled through lexical rules where one word selects another word or preposition, and that word's semantic value is determined by this selected element. It will be shown here how this is accommodated in LFG by two examples: adjective noun constructions, and prepositional verbs.

When a noun is modified by an adjective, it usually allows for genitive nouns or pronouns to come in between, even if the expression is highly non-compositional, as shown in (29).

(29.a) دراجة نارية
darrajah nariyah
bike fiery (motorbike)

(29.b) رأيت دراجة الولد الصغير النارية
ra'itu darrajah al-walad al-saghir al-nariyah
saw I bike the-boy the-young the-fiery
(I saw the young boy's motorbike)

(29.c) C-Structure of the object NP in sentence (29.b)

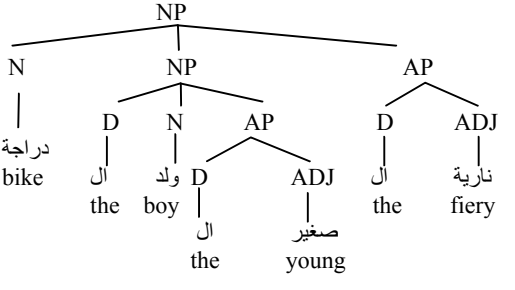

Fig. 3. C-structure of an MWE NP

(29.d) F-Structure of the object NP in sentence (29.b)

```
OBJ   [PRED 'دراجة[bike]'
       MOD      [PRED ' ولد[boy]'
                 DEF +, GEND masc, NUM sg, PERS 3
                 ADJUNCT  [PRED 'صغير [young]'
                           DEF +, GEND masc,
                           NUM sg, PERS 3]
                DEF +, GEND masc, NUM sg, PERS 3]
       ADJUNCT [PRED ' ناري[fiery]'
                DEF +, GEND fem, NUM sg, PERS 3]
       DEF +, GEND fem, NUM sg, PERS 3,
       TRANS motorbike]
```

Fig. 4. F-structure of an MWE NP

This is done by allowing the lexical entry of the noun to select its modifier, as shown by the lexical rule in (30).

(30) دراجة[bike] N {(^PRED='دراجة[bike]' (^ ADJUNCT PRED)=c 'ناري[fiery]'
(^ TRANS)=motorbike
| (^PRED='دراجة[bike]' (^ ADJUNCT PRED)~= 'ناري[fiery]'
(^ TRANS)=bike}.

This means that the translation, or the semantic value, of the noun changes according to the value of the adjunct, or the adjectival modifier. The operator "=c" in the rule means "equal", and "~=" means "not equal".

Similarly, prepositional verbs in Arabic allow for subjects to intervene between verbs and objects as shown by the example in (31). This is why they need to be handled in the syntax.

(31.a) اعتمد الولد على البنت
i'tamada al-waladu 'ala al-bint
relied the-boy on the-girl
(The boy relied on the girl)

(31.b) C-Structure

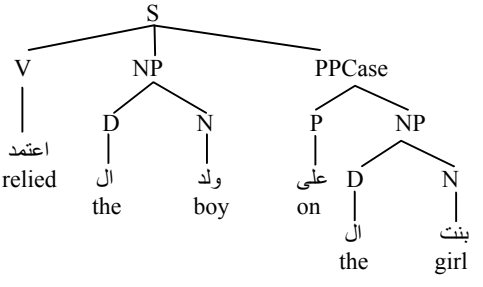

Fig. 5. C-structure of an MWE NP

(31.c) F-Structure

PRED 'اعتمد[rely]<(^ SUBJ)(^ OBJ)'
SUBJ [PRED 'ولد[boy]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND masc, NUM sg, PERS 3
OBJ [PRED 'بنت[girl]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND fem, NUM sg, PERS 3,
PFORM على[on]

F-structure of an MWE NP

Fig. 6. F-structure of an MWE NP

In the c-structure the prepositional verbs looks like a verb followed by a PP. In the f-structure, however, the PP functions as the object of the verb. The semantic value, or PRED, of the preposition is removed. The preposition functions only as a case

assigner and a feature marker to the main object, but it does not subcategorize for an object itself as shown in (32).

(32) على[on] P (^ PFORM)=على[on] (^ PCASE)=gen.

The lexical entry of the verb, as shown in (33), states that the verb subcategorizes for an object with a certain value for the PFORM feature. This means that the object must be preceded by a specified preposition.

(33) اعتمد[rely] V (^ PRED)='اعتمد[rely]<(^ SUBJ)
(^ OBJ)>' (^ OBJ PFORM)=c على[on].

5 Conclusion

The important lesson of this analysis of MWEs is that they must be integrated in the processing and preprocessing stages if we want to obtain any viable linguistic analysis. When MWEs are properly dealt with, they reduce parse ambiguities and give a noticeable degree of certitude to the analysis and machine translation output. This paper explains different types of MWEs and shows what type can be analyzed at what stage.

References

Attia, Mohammed. 2005. Developing Robust Arabic Morphological Transducer Using Finite State Technology. Paper presented at *The 8th Annual CLUK Research Colloquium*, Manchester, UK.

Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki, and Widdows, Dominic. 2003. An Empirical Model of Multiword Expression Decomposability. Paper presented at *the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

Baldwin, Timothy. 2004. Multiword Expressions, an Advanced Course. Paper presented at *The Australasian Language Technology Summer School (ALTSS 2004)*, Sydney, Australia.

Beesley, K. R., and Karttunen, L. 2003. *Finite State Morphology*. Stanford, Calif.: CSLI Publications.

Bentivogli, L., and Pianta, E. 2003. Beyond Lexical Units: Enriching WordNets with Phrasets. Paper presented at *EACL-03*, Budapest, Hungary.

Brun, Caroline. 1998. Terminology finite-state preprocessing for computational LFG. Paper presented at *The 36th conference on Association for Computational Linguistics*, Montreal, Quebec, Canada.

Butt, Miriam, King, Tracy Holloway, Nino, Maria-Eugenia, and Segond, Frederique. 1999. *A Grammar Writer's Cookbook*. Stanford, CA: CSLI.

Calzolari, N., Lenci, A., and Quochi, V. 2002. Towards Multiword and Multilingual Lexicons: between Theory and Practice. Paper presented at *LP2002*, Urayasu, Japan.

Dipper, Stefanie. 2003. Implementing and Documenting Large-Scale Grammars -- German LFG, Institut f˙ur maschinelle Sprachverarbeitung, Institut f˙ur maschinelle Sprachverarbeitung der Stuttgart University: Ph.D.

Fellbaum, Christine ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Guenthner, Frantz, and Blanco, Xavier. 2004. Multi-Lexemic Expressions: an overview. In *Lexique, Syntaxe et Lexique-Grammaire*, eds. Christian Leclère, Éric Laporte, Mireille Piot and Max Silberztein. Philadelphia PA, USA: John Benjamins.

Karttunen, Lauri, Chanod, Jean-Pierre, Grefenstette, G., and Schiller, A. 1996. Regular expressions for language engineering. *Natural Language Engineering* 2:305-328.

Li, W., Zhang, X., Niu, C., Jiang, Y., and Srihari, R. K. 2003. An Expert Lexicon Approach to Identifying English Phrasal Verbs. Paper presented at *The Association for Computational Linguistics (ACL- 2003)*, Sapporo, Japan.

Nerima, Luka, Seretan, Violeta, and Wehrli, Eric. 2003. Creating a Multilingual Collocations Dictionary from Large Text Corpora. Paper presented at *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, Budapest, Hungary.

Oflazer, Kemal, Uglu, Özlem Çetino, and Say, Bilge. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. Paper presented at *Second ACL Workshop on Multiword Expressions: Integrating Processing*, Spain.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan. 2001. Multi-word Expressions: A Pain in the Neck for NLP. Paper presented at *LinGO Working Papers*, Stanford University, CA.

Smadja, Frank, McKeown, Kathleen R., and Hatzivassiloglou, Vasileios. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics* 22:1-38.

Venkatapathy, Sriram. 2004. Overview of my work on Multi-word expressions and Semantic Role Labeling.

Volk, Martin. 1998. The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems. In *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state of the art*, ed. Nico Weber. St. Augustin: gardez-Verlag.