Integrating Morphology with Multi-word Expression Processing in Turkish

Kemal Oflazer and Özlem Çetinoğlu
Human Language and Speech Technology Laboratory
Sabancı University
Istanbul, Turkey
{oflazer,ozlemc}@sabanciuniv.edu

Bilge Say
Informatics Institute
Middle East Technical University
Ankara, Turkey
bsay@ii.metu.edu.tr

Abstract

This paper describes a multi-word expression processor for preprocessing Turkish text for various language engineering applications. In addition to the fairly standard set of lexicalized collocations and multi-word expressions such as named-entities, Turkish uses a quite wide range of semi-lexicalized and non-lexicalized collocations. After an overview of relevant aspects of Turkish, we present a description of the multi-word expressions we handle. We then summarize the computational setting in which we employ a series of components for tokenization, morphological analysis, and multi-word expression extraction. We finally present results from runs over a large corpus and a small gold-standard corpus.

1 Introduction

Multi-word expression extraction is an important component in language processing that aims to identify segments of input text where the syntactic structure and the semantics of a sequence of words (possibly not contiguous) are usually not compositional. Idiomatic forms, support verbs, verbs with specific particle or pre/post-position uses, morphological derivations via partial or full word duplications are some examples of multi-word expressions. Further, expressions such as time-date expressions or proper nouns which can be described with simple (usually finite state) grammars, and whose internal structure is of no real importance to the overall analysis of the sentence, can also be considered under this heading. Marking multi-word expressions in text usually reduces (though not significantly) the number of actual tokens that further processing modules use as input, although this reduction may depend on the domain the text comes from. It can also reduce the multiplicative ambiguity as morphological interpretations of tokens are reduced when they are coalesced into multi-word expressions with usually a single interpretation.

Turkish presents some interesting issues for multi-word expression processing as it makes substantial use of support verbs with lexicalized direct or oblique objects subject to various morphological constraints. It also uses partial and full reduplication of forms of various parts-of-speech, across their whole domain to form what we call *non-lexicalized* collocations, where it is the duplication and contrast of certain morphological patterns that signal a collocation rather than the specific root words used.

In this paper, we describe a multi-word expression processor for preprocessing Turkish text for various language engineering applications. In the next section after a very short overview of relevant aspects of Turkish, we present a rather comprehensive description of the multi-word expressions we handle. We then summarize the structure of the multi-word expression processor which employs a series of components for tokenization, morphological analysis, conservative non-statistical morphological disambiguation, and multi-word expression extraction. We finally present results from runs over a large corpus and a small gold-standard corpus.

1.1 Related Work

Recent work on multi-word expression extraction, use three basic approaches: statistical, rule-based, and hybrid. Statistical approaches require a corpus that contains significant numbers of occurrences of multi-word expressions. But even if the corpus consists of millions of words, usually, the frequencies of multi-word expressions are too low for statistical extraction. Baldwin and Villavicencio (2002) indicate that "two-thirds of verb-particle constructions occur at most three times in the overall corpus, meaning that any extraction method must be able to handle extremely sparse data." They use a rule-based method to extract multi-word expressions in the form of a head verb and a single obligatory preposition employing a tagger augmented with an

existing chunking system with which they first identify the particle chunked and then turn back for the verb part of the construction.

Piao et al. (2003) employ their semantic field annotator USAS, containing 37,000 words and a template list of 16,000 multi-word units, all constructed manually from various resources, in order to extract multi-word expressions. The evaluation indicates a high precision (over 90%) but the estimated recall is about 40%. Deeper investigation on the corpus has indicated that two-thirds of the multi-word expressions occur in the corpus once or twice, verifying the fact that the statistical methods filtering low frequencies would fail.

Urizar et al. (2000) describe a Basque terminology extraction system which covered multi-word term extraction as a subset. As Basque is a highly inflected agglutinative language like Turkish, morphological information is exploited to better define multi-word patterns. Their lemmatizer/tagger EU-SLEM, consists of a tokenizer followed by two subsystems for the treatment of single word and multi-word expressions, and a disambiguator. The proposed term extraction tool uses the tagged input as the input of a shallow parsing phase which consists of regular expressions representing morphosyntactic patterns. The final step uses statistical measures to eliminate incorrect candidates.

The basic disadvantages of rule-based approaches are that they usually lack flexibility, and it is a time-consuming and never ending process to try to cover a high percentage of the multi-word expressions in a language with rules and predefined lists. The LINGO group which defines multi-word expressions as "a pain in the neck for NLP" (Sag et al., 2002), suggests hybrid approaches using rule based approaches to identify possible multi-word expressions out of a corpus and using statistical methods to enhance the results obtained.

2 Multi-word expressions in Turkish

Turkish is an Ural-Altaic language, having agglutinative word structures with productive inflectional and derivational processes. Most derivational phenomena take place within a word form, but there are certain derivations involving partial or full reduplications that are best considered under the notion of multi-word expressions.

Turkish word forms consist of morphemes concatenated to a root morpheme or to other morphemes,

much like beads on a string. Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various morphophonemic processes such as vowel harmony, vowel and consonant elisions. The morphotactics of word forms can be quite complex when multiple derivations are involved. For instance, the derived modifier sağlamlaştırdığımızdaki[1] would be represented as:[2]

```
sağlam+Adj
    ^DB+Verb+Become
    ^DB+Verb+Caus+Pos
    ^DB+Adj+PastPart+P1sg
    ^DB+Noun+Zero+A3sg+Pnon+Loc
    ^DB+Adj
```

This word starts out with an adjective root and after five derivations, ends up with the final part-of-speech adjective which determines its role in the sentence.

Turkish employs multi-word expressions in essentially four different forms:

1. *Lexicalized Collocations* where all components of the collocations are fixed,

2. *Semi-lexicalized Collocations* where some components of the collocation are fixed and some can vary via inflectional and derivational morphology processes and the (lexical) semantics of the collocation is not compositional,

3. *Non-lexicalized Collocations* where the collocation is mediated by a morphosyntactic pattern of duplicated and/or contrasting components – hence the name *non-lexicalized*, and

4. *Multi-word Named-entities* which are multi-word proper names for persons, organizations, places, etc.

2.1 Lexicalized Collocations

Under the notion of lexicalized collocations, we consider the usual fixed multi-word expressions

---

[1]Literally, "(the thing existing) at the time we caused (something) to become strong". Obviously this is not a word that one would use everyday. Turkish words (excluding noninflecting frequent words such as conjunctions, clitics, etc.) found in typical text average about 10 letters in length.

[2]Please refer to the list of morphological features given in Appendix A for the semantics of some of the non-obvious symbols used here.

whose resulting syntactic function and semantics are not readily predictable from the structure and the morphological properties of the constituents.

Here are some examples of the multi-word expressions that we consider under this grouping:[3,4]

(1) *hiç olmazsa*

- `hiç(never)+Adverb`
  `ol(be)+Verb+Neg+Aor+Cond+A3sg`
- `hiç_olmazsa+Adverb`
  "at least" (literally "if it never is")

(2) *ipe sapa gelmez*

- `ip(rope)+Noun+A3sg+Pnon+Dat`
  `sap(handle)+Noun+A3sg+Pnon+Dat`
  `gel(come)+Verb+Neg+Aor+A3sg`
- `ipe_sapa_gelmez+Adj`
  "worthless" (literally "(he) does not come to rope and handle")

2.2 Semi-lexicalized Collocations

Multi-word expressions that are considered under this heading are compound and support verb formations where there are two or more lexical items the last of which is a verb or is a derivation involving a verb. These are formed by a lexically adjacent, direct or oblique object, and a verb, which for the purposes of syntactic analysis, may be considered as single lexical item: e.g., *saygı dur-* (literally *to stand (in) respect* – to pay respect), *kafayı ye-* (literally *to eat the head* – to get mentally deranged), etc.[5] Even though the other components can themselves be inflected, they can be assumed to be fixed for the purposes of the collocation, and the collocation assumes its morphosyntactic features from the last verb which itself may undergo any morphological derivation or inflection process. For instance in

(3) *kafayı ye-*

- `kafa(head)+Noun+A3sg+Pnon+Acc`
  `ye(eat)+Verb...`

---

[3] In every group we first list the morphological features of all the tokens, one on every line (with the glosses for the roots), and then provide the morphological features of the multi-word construct and then provide glosses and literal meanings.

[4] Please refer to the list of morphological features given in Appendix A for the semantics of some of the non-obvious symbols used here.

[5] Here we just show the roots of the verb with - denoting the rest of the suffixes for any inflectional and derivational markers.

- `kafayı_ye+Verb...`
  "get mentally deranged" (literally "eat the head")

the first part of the collocation, the accusative marked noun *kafayı*, is the fixed part and the part starting with the verb *ye-* is the variable part which may be inflected and/or derived in myriads of ways. For example the following are some possible forms of the collocation:

- kafayı ye*dim* "I got mentally deranged"
- kafayı yi*yeceklerdi* "they were about to get mentally deranged"
- kafayı yi*yenler* "those who got mentally deranged"
- kafayı ye*diği* "the fact that (s/he) got mentally deranged"

Under certain circumstances, the "fixed" part may actually vary in a rather controlled manner subject to certain morphosyntactic constraints, as in the idiomatic verb:

(4) *kafa(yı) çek-*

- `kafa(head)+Noun+A3sg+Pnon+Acc`
  `çek(pull)+Verb...`
- `kafa_çek+Verb...`
  "consume alcohol" (but literally "to pull the head")

(5) *kafaları çek-*

- `kafa+Noun+`**`A3pl`**`+Pnon+Acc`
  `çek+Verb...`
- `kafa_çek+Verb...`
  "consume alcohol" (but literally "to pull the heads")

where the fixed part can be in the nominative or the accusative case, and if it is in the accusative case, it may be marked plural, in which case the verb has to have some kind of plural agreement (i.e., first, second or third person plural), *but no possessive agreement markers are allowed.*

In their simplest forms, it is sufficient to recognize a sequence of tokens one of whose morphological analyses matches the corresponding pattern, and then coalesce these into a single multi-word expression representation. However, some or all variants of these and similar semi-lexicalized collocations

present further complications brought about by the relative freeness of the constituent order in Turkish, and by the interaction of various clitics with such collocations.[6]

When such multi-word expressions are coalesced into a single morphological entity, the ambiguity in morphological interpretation is reduced as we see in the following example:

(6) *devam etti*

- ```
  devam(continuation)+Noun+A3sg
                          +Pnon+Nom
  *deva(therapy)+Noun+A3sg+P1sg+Nom

  et(make)+Verb+Pos+Past+A3sg
  *et(meat)+Noun+A3sg+Pnon+Nom
                ^DB+Verb+Past+A3sg
  ```
- ```
  devam_et+Verb+Pos+Past+A3sg
  ```
  "(he) continued" (literally "made a continuation")

Here, when this semi-lexicalized collocation is recognized, other morphological interpretations of the components (marked with a `*` above) can safely be removed, contributing to overall morphological ambiguity reduction.

2.3 Non-lexicalized Collocations

Turkish employs quite a number of non-lexicalized collocations where the sentential role of the collocation has (almost) nothing to do with the parts-of-speech and the morphological features of the individual forms involved. Almost all of these collocations involve partial or full duplications of the forms involved and can actually be viewed as morphological derivational processes mediated by reduplication across multiple tokens.

The *morphological feature representations* of such multi-word expressions follow one of the patterns:

1) $\omega \ \omega$ 3) $\omega + X \ \omega + Y$
2) $\omega \ Z \ \omega$,   4) $\omega_1 + X \ \omega_2 + X$

where ω is the duplicated string comprising the root, its part-of-speech and possibly some additional morphological features encoded by any suffixes. X and Y are further duplicated or contrasted morphological patterns and Z is a certain clitic token. In

---

[6]The question and the emphasis clitics which are written as separate tokens, can occasionally intervene between the components of a semi-lexicalized collocation. We omit the details of these due to space restrictions.

duplications of type 4, it is possible that ω_1 is different from ω_2.

Below we present list of the more interesting non-lexicalized expressions along with some examples and issues.

- When a noun appears in duplicate following the first pattern above, the collocation behaves like a manner adverb, modifying a verb usually to the right. Although this pattern does not necessarily occur with every possible noun, it may occur with many (countable) nouns without much of a further semantic restriction. Such a sequence has to be coalesced into a representation indicating this derivational process as we see below.

(7) *ev ev* ($\omega \ \omega$)

- ```
  ev(house)+Noun+A3sg+Pnon+Nom
  ev+Noun+A3sg+Pnon+Nom
  ```
- ```
  ev+Noun+A3sg+Pnon+Nom^DB+Adverb+By
  ```
  "house by house" (literally "house house")

- When an adjective appears in duplicate, the collocation behaves like a manner adverb (with the semantics of *-ly* adverbs in English), modifying a verb usually to the right. Thus such a sequence has to be coalesced into a representation indicating this derivational process.

(8) *yavaş yavaş* ($\omega \ \omega$)

- ```
  yavaş(slow)+Adj
  yavaş+Adj
  ```
- ```
  yavaş+Adj^DB+Adverb+Ly
  ```
  "slowly" (literally "slow slow")

This kind of duplication can also occur when the adjective is a derived adjective as in

(9) *hızlı hızlı* ($\omega \ \omega$)

- ```
  hız(speed)+Noun+A3sg+Pnon+Nom
              ^DB+Adj+With
  hız+Noun+A3sg+Pnon+Nom
              ^DB+Adj+With
  ```
- ```
  hız+Noun+A3sg+Pnon+Nom
      ^DB+Adj+With^DB+Adverb+Ly
  ```
  "rapidly" (literally "with-speed with-speed")

- Turkish has a fairly large set of onomatopoeic words which always appear in duplicate and function as manner adverbs. The words by themselves have no other usage and literal meaning, and mildly resemble sounds produced by natural or artificial objects. In these cases, the root word almost always is reduplicated but need not be, but both words should be of the part-of-speech category `+Dup` that we use to mark such roots.

(10) *harıl hurul* ($\omega_1 + X\ \omega_2 + X$)

- `harıl+Dup`
  `hurul+Dup`
- `harıl_hurul+Adverb+Resemble`
  "making rough noises" (no literal meaning)

- Duplicated verbs with optative mood and third person singular agreement function as manner adverbs, indicating that another verb is executed in a manner indicated by the duplicated verb:

(11) *koşa koşa* ($\omega\ \omega$)

- `koş(run)+Verb+Pos+Opt+A3sg`
  `koş(run)+Verb+Pos+Opt+A3sg`
- `koş+Verb+Pos+^DB+Adverb+ByDoingSo`
  "by running" (literally "let him run let him run")

- Duplicated verbs in aorist mood with third person agreement and first positive then negative polarity, function as temporal adverbs with the semantics "as soon as one has *verb*ed"

(12) *uyur uyumaz* ($\omega + X\ \omega + Y$)

- `uyu+Verb+`**`Pos+Aor+A3sg`**
  `uyu+Verb+`**`Neg+Aor+A3sg`**
- `uyu+Verb+Pos+^DB+Adverb+AsSoonAs`
  "as soon as (he) sleeps" (literally "(he) sleeps (he) does not sleep")

It should be noted that for most of the non-lexicalized collocations involving verbs (like (11) and (12) above), the verbal portion before the inflectional marking mood can have additional derivational markers and all such markers have to duplicate.

(13) *sağlamlaştırır sağlamlaştırmaz* ($\omega + X\ \omega + Y$)

- `sağlam+Adj^DB+Verb+Become`
  `^DB+Verb+Caus^DB+Verb+`**`Pos+Aor+A3sg`**
  `sağlam+Adj^DB+Verb+Become`
  `^DB+Verb+Caus^DB+Verb+`**`Neg+Aor+A3sg`**

- `sağlam+Adj^DB+Verb+Become+`
  `^DB+Verb+Caus+Pos`
  `^DB+Adverb+AsSoonAs`
  "as soon as (he) fortifies (causes to become strong)"

Another interesting point is that non-lexicalized collocations can interact with semi-lexicalized collocations since they both usually involve verbs. For instance, when the verb of the semi-lexicalized collocation example in (5) is duplicated in the form of the non-lexicalized collocation in (12), we get

(14) *kafaları çeker çekmez*

In this case, first the non-lexicalized collocation has to be coalesced into

(15) *kafaları* `çek+Verb+Pos`
`^DB+Adverb+AsSoonAs`

and then the semi-lexicalized collocation kicks in, to give

(16) `kafa_çek+Verb+Pos`
`^DB+Adverb+AsSoonAs`
("as soon as (we/you/they) get drunk")

Finally, the following non-lexicalized collocation involving adjectival forms involving duplication and a question clitic is an example of the last type of non-lexicalized collocation.

(17) *güzel mi güzel* ($\omega\ Z\ \omega$)

- `güzel+Adj`
  `mi+Ques`
  `güzel+Adj`

- `güzel+Adj+Very`
  "very beautiful" (literally "beautiful (is it?) beautiful")

2.4 Named-entities

Another class of multi-word expressions that we process is the class of multi-word named-entities denoting persons, organizations and locations. We essentially treat these just like the semi-lexicalized collocation discussed earlier, in that, when such named-entities are used in text, all but the last component are fixed and the last component will usually undergo certain morphological processes demanded by the syntactic context as in
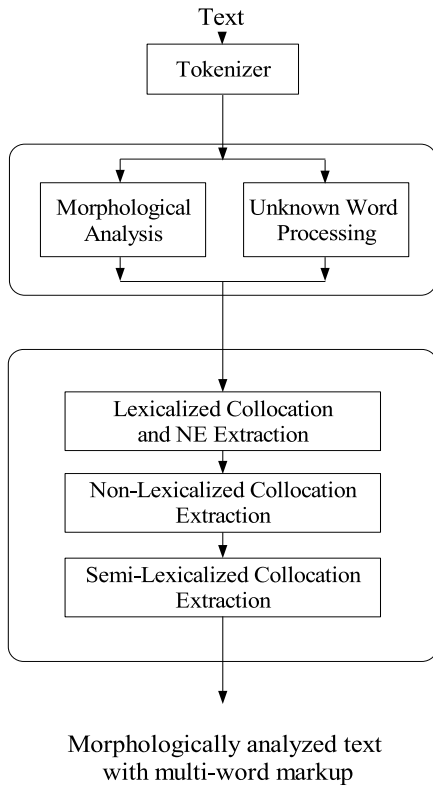
Text

Figure 1: The architecture of the multi-word expression extraction processor

(18) *Türkiye Büyük Millet Meclisi'nde ....*[7]

Here, the last component is case marked and this represents a case marking on the whole named-entity. We package this as

(19) `Türkiye_Büyük_Millet_Meclisi`
`    +Noun+Prop+A3sg+Pnon+Loc`

To recognize these named entities we use a rather simple approach employing a rather extensive database of person, organization and place names, developed in the context of a previous project, instead of using a more sophisticated named-entity extraction scheme.

3 The Structure of the Multi-word Expression Processor

Our multi-word expression processor is a multi-stage system as depicted in Figure 1. The first component is a standard tokenizer which splits input text into constituent tokens. These then go into

[7]In the Turkish Grand National Assembly.

a wide-coverage morphological analyzer (Oflazer, 1994) implemented using Xerox finite state technology (Karttunen et al., 1997), which generates, for all tokens, all possible morphological analyses. This module also performs unknown processing by postulating possible noun roots and then trying to parse the rest of a word as a sequence of possible Turkish suffixes. The morphological analysis stage also performs a very conservative non-statistical morphological disambiguation to remove some very unlikely parses based on unambiguous contexts. Figure 2 shows a sample Turkish text that comes out of morphological processing, about to go into multi-word expression extraction.

```
Kistin       kist+Noun+A3sg+P2sg+Nom
             kist+Noun+A3sg+Pnon+Gen
sağlığımı    sağlık+Noun+A3sg+P1sg+Acc
             sağ+AdjˆDB+Noun+Ness+
                 A3sg+P1sg+Acc
sıkıntıya    sıkıntı+Noun+A3sg+Pnon+Dat
sokacak      sok+Verb+Pos+Fut+A3sg
             sok+Verb+PosˆDB+Adj
                 +FutPart+Pnon
herhangi     herhangi+Adj
bir          bir+Det
             bir+Num+Card
             bir+Adj
             bir+Adverb
etkisi       etki+Noun+A3sg+P3sg+Nom
söz          söz+Noun+A3sg+Pnon+Nom
konusu       konu+Noun+A3sg+P3sg+Nom
değil        değ+VerbˆDB+Verb+Pass
                 +Pos+Imp+A2sg
             değil+Conj
             değil+Verb+Pres+A3sg
.            .+Punc
```

Figure 2: Output of the morphological analyzer

The multi-word expression extraction processor has three stages with the output of one stage feeding into the next stage:

1. The first stage handles lexicalized collocations and multi-word named entities.

2. The second stage handles non-lexicalized collocations.

3. The third stage handles semi-lexicalized collocations. The reason semi-lexicalized collocations are handled last, is that any duplicate verb formations have to be processed before compound verbs are combined with their lexicalized complements (cf. examples (14) – (16) above).

The output of the multi-word expression extraction processor for the relevant segments in Figure 2 is given in Figure 3.

The multi-word expression extraction processor has been implemented in Perl. The rule bases for the three stages are maintained separately and then compiled offline into regular expressions which are then used by Perl at runtime.

```
...
sıkıntıya_sokacak   sıkıntıya_sok+Verb
                    +Pos+Fut+A3sg
                    sıkıntıya_sok+Verb
                    +Pos^DB+Adj
                    +FutPart+Pnon
herhangi_bir        herhangi_bir+Det
...
söz_konusu          söz_konu+Noun+A3sg
                    +P3sg+Nom
...
```

Figure 3: Output of the multi-word expression extraction processor

Table 1 presents statistics on the current rule base of our multi-word expression extraction processor: For named entity recognition, we use a list of about

Rule Type	Number of Rules
Lexicalized Colloc.	363
Semi-lexicalized Colloc.	731
Non-lexicalized Colloc.	16

Table 1: Rules base statistics

60,000 first and last names, a list of about 16,000 multi-word organization and place names.

4 Evaluation

To improve and evaluate our multi-word expression extraction processor, we used two corpora of news text. We used a corpus of about 730,000 tokens to incrementally test and improve our semi-lexicalized rule base, by searching for compound verb formations, etc. Once such rules were extracted, we tested our processor on this corpus, and on a small corpus of about 4200 words to measure precision and recall. Table 2 provides some statistics on these corpora.

Table 3 shows the result of multi-word expression extraction on the large (training) corpus. It should be noted that we only mark multi-word named-entities, not all. Thus many references to persons by

Corpus	Number of Tokens	Avg. Analyses per Token
Large Corpus	729,955	1.760
Small Corpus	4,242	1.702

Table 2: Corpora Statistics

their last name are not marked, hence the low number of named-entities extracted.[8] As a result of

MW Type	Number Extracted
Lexicalized Colloc.	3,883
Semi-lexicalized Colloc.	9,173
Non-lexicalized Colloc.	220
Named-Entities	4,480
Total	17,750

Table 3: Multi-word expression extraction statistics on the large corpus

this extraction, the average number of morphological parses per token go from 1.760 down to 1.745.

Table 4 shows the result of multi-word expression extraction on the small corpus. We also manu-

MW Type	Number Extracted
Lexicalized Colloc.	15
Semi-lexicalized Colloc.	62
Non-lexicalized Colloc.	0
Named-Entities	99
Total	176

Table 4: Multi-word expression extraction statistics on the small corpus

ally marked up the small corpus into a gold-standard corpus to test precision and recall. The results in Table 4 correspond to an overall recall of 65.2% and a precision of 98.9%, over all classes of multi-word expressions. When we consider all classes except named-entities, we have a recall of 60.1% and a precision of 100%. An analysis of the errors and missed multi-word expressions indicates that the test corpus had a certain variant of a compound verb construction that we had failed to extract from the larger corpus we used for compiling rules. Failing to extract the multi-word expressions for that compound verb accounted for most of the drop in recall. Since we are currently using a rather naive named-entity extraction scheme,[9] re-

[8]Since this is a very large corpus, we have no easy way of obtaining accurate precision and recall figures.

[9]As opposed to a general purpose statistical NE extractor that we have developed earlier (Tür et al., 2003).

call is rather low as there are quite a number of foreign multi-word named-entities (persons and organizations mostly) that do not exist in our database of named-entities. On the other hand, since named-entity extraction for English is a relatively mature technology, we can easily integrate an existing tool to improve our recall.

5 Conclusions

This paper has described a multi-word expression extraction system for Turkish for handling various types of multi-word expressions such as semi-lexicalized and non-lexicalized collocations which depend on the recognition of certain morphological patterns across tokens. Our results indicate that with about 1100 rules (most of which were extracted from a large "training corpus" searching for patterns involving a certain small set of support verbs), we were able get almost 100% precision and around 60% recall on a small "test" corpus. We expect that with additional rules from dictionaries and other sources we will improve recall significantly.

6 Acknowledgments

We thank Orhan Bilgin for helping us compile the multi-word expressions.

References

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL 2002)*, pages 99–105.

Lauri Karttunen, Tamas Gaal, and Andre Kempe. 1997. Xerox Finite-State Tool. Technical report, Xerox Research Centre Europe.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 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)*, pages 1–15.

Gökhan Tür, Dilek Zeynep Hakkani-Tür, and Kemal Oflazer. 2003. A statistical information extraction systems for Turkish. *Natural Language Engineering*, 9(2).

R. Urizar, N. Ezeiza, and I. Alegria. 2000. Morphosyntactic structure of terms in Basque for automatic terminology extraction. In *Proceedings of the ninth EURALEX International Congress*.

A Morphosyntactic Features For Turkish

This section lists the features and their semantics for the morphological representations used in the text. `^DB` marks a derivation boundary.

- Parts-of-speech: `+Noun`, `+Adjective`, `+Adverb`, `+Verb`, `+Dup` (for onomatopoeic words which always appear in duplicate), `+Question` (yes/no question marker clitic), `+Number`, `+Determiner`

- Agreement: `+A[1-3][sg-pl]`, e.g., `+A3pl`.

- Possessive agreement: `+P[1-3][sg-pl]` and `+Pnon`, e.g., `+P1sg`

- Case: `+Nominative`, `+Accusative`, `+Locative`, `+Ablative`, `+Instrumental`, `+Genitive`, `+Dative`.

- Miscellaneous Verbal Features: `+Causative`, `+Passive`, `+Positive` Polarity, `+Negative` Polarity, `+Optative` Mood, `+Aorist` Aspect, `+Become`, `+Conditional` Mood, `+Imperative` Mood, `+Past` tense

- Miscellaneous POS Subtypes: Adverbs: `+By` (as in "house by house"), `+ByDoingSo`, (as in "he came by running"), `+Resemble` (as in "he made sounds resembling .."), `+Ly` (as in "slowly") `+AsSoonAs` (as in "he came down as soon as he woke up"); Adjectives: `+With` (as in "the book with red cover"), `+FutPart` – future participle – as in ("the boy who will come"); Nouns: `+Proper` Noun, `+Ness` (as in "sick-ness"), `+FutPart` – future participle fact – as in ("I know that he will come") ; Numbers: `+Cardinal`