Finite State Analysis of Prepositional Phrases in Maltese

Michael Rosner

Department of Computer Science and AI,
University of Malta

Abstract. We address the general problem faced by designers of computational lexica: that of relating *surface* forms to underlying *lexical* forms through the vehicle of a precise linguistic description expressed in a suitable formalism. For such a description to be useful, it must be possible for the relation to be *computable*: given a surface element, we need to compute the corresponding lexical element or elements, and vice cersa. Below we concentrate upon the description of a minimal example: prepositional phrases, a reasonably well-defined and compact subset of Maltese surface phenomena that exemplifies many of the difficulties that are typical of Semitic languages.

1 Introduction

The work reported here is carried out under the general umbrella of Maltilex, a research programme supported at the University of Malta that aims to develop algorithmic and data resources for the Maltese Language (see Rosner et. al [8, 9]). Most of the effort so far has been directed towards the goal of automatically structuring lexical entries acquired from corpora. LST is the name given to the structuring technique, which is the subject of a recent MSc dissertation by Angelo Dalli (see [5] for a detailed descripion of the technique).

LST is essentially a statistical process that has no knowledge of the internal structure of words. This is both a blessing and a curse. The great advantage is that linguistic expertise is not necessary. The disadvantage is that it if you are in possession of linguistic expertise, it may be difficult or impossible to incorporate it into the system.

Another major goal of the project is therefore to find ways of incoporating certain kinds of linguistic knowledge (morpholotactic knowledge in particular). For this to be possible we must (a) improve our understanding of morpho-grammatical aspects of the language (b) construct computational models and (c) develop means of biasing LST to take account of them.

This draft represents a first step in that direction and presents a description that deals with a well-defined part of the problem of mapping between underlying lexical strings and orthographic words (i.e. surface strings appearing between spaces).

Here we restrict our attention to simple prepositional phrases which are made up of a preposition, an (optional) article, and a noun. The first part of the abstract presents the linguistic facts. This is followed by a computable description using xfst [3], a finite state notation and compiler that has been developed at Xerox, and a brief presentation of the grammar.

We hope to extend this draft into a full paper in due course.

2 The Maltese Language Alphabet

The Maltese alphabet comprises 24 consonants and 6 vowels as shown below:

consonants	**vowels**
b c d f	*a*
ġ g għ h	*e*
ħ j k l	*i*
m n p q	*ie*
r s t v	*o*
w x ż z	*u*

2.1 The Definite Article in Maltese

Definiteness in Maltese is expressed by preposing the definite article to the noun to form a single word. The unmarked form of the article takes the orthographic form *l-* (including the dash), e.g. *l-ittra* (the letter), but this can change depending on the surrounding morpho-phonological environment.

The two main phonological phenomena giving rise to these changes are referred to (see [6] as /i/-epenthesis and consonant assimilation.

/i/-epenthesis involves insertion of the vowel *i* and serves to aid pronunciation. When this occurs purely as a result of characteristics of the noun alone, it is called *inner* epenthesis, whilst if it concerns characteristics of the environment outside the noun, it is *outer* epenthesis.

Inner epenthesis arises when the noun begins with a cluster of consonants beginning with s or x. An i is inserted before the noun. An example of a noun fulfilling these conditions is *skola* (school), for which the definite form is *l-iskola*.

Outer epenthesis occurs when the noun begins with a consonant and the article is not preceded by a word that ends in a vowel. An example of such a word is *tak* (he-gave-you). Hence *tak il-ktieb/il-karta* (he-gave-you the book/the paper). This is to be contrasted with *tani l-ktieb/l-karta* (you-gave-me the book/the paper). Note that outer epenthesis also occurs when the article is at the beginning of string. So when stanfing alone, we say *il-karta*.

Consonant assimilation takes place whenever the initial consonant of the noun is one the so-called "sun letters": ċ, d, s, r, t, ż, x. In these circumstances, we write *ix-xemx* (the sun), *id-dar* (the house) rather than *il-xemx* or *il-dar*.

2.2 Prepositions

Maltese prepositions that demonstrate interesting morphological behaviour[1] are shown below together with their nearest English equivalents:

[1] there are several others that do not

Maltese	English
ma'	with
ta'	of
sa	to
ġo	in(to)
bi	with
fi	in
lil	to
għal	for
bħal	like
minn	from

These forms are used when the preposition immediately precedes a noun without an article. This can be for a variety of reasons, e.g. (i) the noun is proper and doesn't take an article (e.g. *lil Mike* - to Mike), (ii) the noun is indefinite (e.g. *bħal għalliem* - like a teacher), (iii) the noun is definite in virtue of something other than the article (*ta' ommi* of my mother, where the possessive is formed with the suffix *i*).

There are some exceptions, however: *i* of *bi* and *fi* is replaced with apostrophe if the noun begins with a vowel, or with only one consonant; the result is joined to the next word as illustrated by *b'ommi* (with my mother), *f'Malta* (in Malta), *b'tifel* (with a boy).

The vowels of *ma'* and *ta'* are dropped before a word beginning with a vowel. Hence *m'ommi, t'Anna* (with my mother, of Anna).

2.3 Preposition + Article

When a preposition is immediately followed by the definite article, it is joined with it to form one word. An example is *sal-Belt* (to the city), which is formed by joining the preposition *sa* (to), the article and the noun. Notice that the result is still a single word.

An exception to this rule is the preposition *minn*, which, when joined with the article in this way, becomes *mill-*. However (see also below), before a noun beginning with *l*, one of the *l*s is dropped: *mil-Lebanon* (from Lebanon).

The prepositions *bi* (with) and *fi* (in) also follow this rule, e.g. *fil-forn* (in the oven), *bil-karozza* (with the car). However, if the noun begins with a vowel, the *i* is dropped, whether or not the vowel is the result of inner epenthesis. Hence we have *fl-iskola* (in the school).

Prepositions ending in *l* (*bħal*, *għal*, *lil*) also have special behaviour. With nouns beginning with letters other than *l*, they behave normally, so we can have *għall-Karnival* (for Carnival), *bħall-Ingliżi* (like the English). However, whenever the word after the article begins with the consonant *l*, that consonant is dropped (to avoid having three *l*s in succession. Hence we have *għal-lukanda* (for the hotel), *lil-Libjan* (to the Libyan).

Consonant assimilation with the article takes place as before when the noun begins with a sun letter: so we have *fid-dar* not *fil-dar*.

The prepositions *ta'* (of) and *ma'* both include an apostrophe (which these cases stand in for an omitted silent letter *għ*) which is dropped when these prepositions are joined to the article: *tal-bieraħ* (of yesterday), *mat-tifel* (with the boy).

3 Computational Aspects

3.1 xfst

xfst is short for Xerox finite-state tool, and is one of a family of finite-state engines developed at Xerox for defining, exploring, and extending the potential of finite state technologies for language engineering purposes.

3.2 Description

In the description below we make heavy use of xfst's replace operator (see [7]) used both conditionally and unconditionally.

Character Classes and Words We begin with the fundamental character classes

```
define V [a | e | i | o | u | ie];
define C [b | "_c" | d | f | "_g"
         | g | h | "_h" | j | k | l
         | m | n | m | p | q | r | s
         | t | v | w | x | "g_h"
         | "_z" | z ];
```

together with the prepositions mentioned above:

```
define PREP [ {ma'} | {ta'} | {sa} |
              {bi} | {fi} | {minn} |
              {lil} | ["g_h" a l]  |
              [b "_h" a l] | [ "_g" o ] ];
```

A representative set of nouns is included, respectively beginning with a vowel, two consonants, a sun letter, and a consonant cluster starting with s.

```
define NOUN [ {ittra} | {ktieb} |
              {xemx} | {skola}];
```

There are also two verbs respectively ending in a vowel and a consonant that will be used to demonstrate epenthesis.

```
define VERB [ {tak} | {tini} ];
```

3.3 Rules

Conventions "+" is used for morpheme/word boundaries, whilst "ws" stands for whitespace or beginning/end of input string (the latter defined as the regular expression

```
[%+ | .#.]
```

Preposition without Article

```
define bificontract
  [ i %+ -> ’ || ws [b|f] _ [C V|V]];
```

Article The article is represented on the lexical side by the abstract character `L`. We first present the rules that carry out inner and outer epenthesis:

```
define iepenthesis
 [[..] -> i || L %+ _ [s C]];
define oepenthesis
 [[..] -> i || C %+ _ L];
```

which are combined together by composition

```
define epenthesis
 oepenthesis .o. iepenthesis;
```

The surface realisation of the article is governed by the following set of rules. The first converts the morpheme boundary into a “-”

```
define ajoin [%+ -> %- || L _];
```

whilst the second ensures that the abstract `L` respects the behaviour demanded by Sun-letters.

```
define atran
  [ L -> "_c" || _ %- "_c"] .o.
  [ L -> d || _ %- d] .o.
  [ L -> s || _ %- s] .o.
  [ L -> t || _ %- t] .o.
  [ L -> x || _ %- x] .o.
  [ L -> "_z" || _ %- "_z"] .o.
  [ L -> l];
```

Finally the two rules are combined together, again using compostion. This time the order is relevant since the atran rules count on the transduction from “+” to “-”.

```
define art ajoin .o. atran;
```

Preposition with Article The first rule states that if a preposition is followed by an article, they are assimilated, i.e. the morpheme boundary disappears. This is the basic preposition rule:

```
define prepbasic
 [ %+ -> [..] || PREP _ L];
```

The use of `[..]` rather than `0` facilitates using the rule in reverse

Next, the exceptions. For *bi* and *fi* the *i* disappears if the noun ahead of the article starts with a vowel.

```
define bfil2bfl
 [ i -> [..] || ws [b | f] _  L %+ V];
```

The ordering of this rule is delicate, since to operate properly it has to be invoked take place *after* (inner) epenthesis, which could potentially insert a vowel after the article.

A second exception is the behaviour shown by prepositions that end in *l*. As already mentioned, this can be assimilated under a general rule that prevents more than three identical consonants from ever appearing adjacent to each other. Unfortunately it is difficult to state a rule of this degree of generality in xfst: we are forced to address the specific case specific case with a rule like this - and the ordering is still critical.

```
define l32
  [l L -> L || _ %+ l];
```

3.4 Results

The problem under investigation can be regarded thus: given our definitions, we want to define a system that will transduce between strings generated by the expression

```
(VERB %+) PREP (%+ L) %+ NOUN
```

and the underlying lexical representation.

4 Conclusion and Future Work

This article is an incomplete draft of a forthcoming paper that will include examples from an online demonstration, a discussion of the quality of results, and suggestions for extending the analysis to other linguistic phenomena in Maltese. The accompanying talk will address some of these points.

5 Acknowledgements

I should like to thank the University of Malta for supporting the research described in this draft. I also thank my colleagues at in the Maltilex project and particularly Ray Fabri,for their helpful comments. Finally, I am indebted to Ken Beesley for guiding me through some of the less intuitive features of xfst.

References

1. Alfred. V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ, 1972.
2. American Psychological Association. *Publications Manual*. American Psychological Association, Washington, DC, 1983.
3. K. Beesley and L. Karttunen. *Finite State Machinery: Tools and Techniques*. Cambridge University Press, Cambridge, forthcoming.
4. Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133, 1981.
5. Angelo Dalli. *Computational Lexicon for Maltese*. University of Malta, MSc. Dissertation, Dept. CSAI, 2002.
6. Ray Fabri. Definiteness marking and the structure of the np in maltese. *Verbum*, XXIII(2):153–173, 2001.
7. L. Karttunen. The replace operator. In E Roche and Y. Schabes, editors, *Finite State Language Processing*, pages 115–147, 1997.
8. M. Rosner, J. Caruana, and R. Fabri. Maltilex: A computational lexicon for maltese. In M. Rosner, editor, *Computational Approaches to Semitic Languages: Proceedings of the Workshop held at COLING-ACL98, Université de Montréal, Canada*, pages 97–105, 1998.
9. M. Rosner, J. Caruana, and R. Fabri. Linguistic and computational aspects of maltilex. In *Arabic Translation and Localisation Symposium: Proceedings of the Workshop, Tunis*, pages 2–10, 1999.