Maltilex: A Computational Lexicon for Maltese

M. Rosner, J. Caruana and R. Fabri
University of Malta, Msida MSD06, Malta
mros@cs.um.edu.mt, jcar1@um.edu.mt, rfab1@um.edu.mt

Abstract

The project described in this paper, which is still in the preliminary phase, concerns the design and implementation of a computational lexicon for Maltese, a language very much in current use but so far lacking most of the infrastructure required for NLP. One of the main characteristics of Maltese, a source of many difficulties, is that it is an amalgam of different language types (chiefly Semitic and Romance), as illustrated in the first part of the paper. The latter part of the paper describes our general approach to the problem of constructing the lexicon.

1 Introduction

With few exceptions (e.g. Galea (1996)) Maltese is pretty much virgin territory as far as language processing is concerned, and therefore one question worth asking is: where to begin? There are basically two extreme positions that one can adopt in answering this question. One is to attack a variety of applications first, e.g. translation, speech, dialogue etc., and hope that in so doing, enough general expertise can be acquired to build the basis of an NLP culture that is taken for granted with more computationally established languages. The other extreme is to attack the linguistic issues first, since, for whatever reason, there is currently rather little in the way of an accepted linguistic framework from which to design computational materials.

We have decided to adopt the middle ground by embarking upon the construction of a substantial machine-tractable lexicon of the language, since whether we think in terms of applications or linguistic theory, the lexicon is clearly a resource of fundamental importance.

The construction of the lexicon involves two rather separate subtasks which may in practice become interleaved. The first is the identification of a set of lexical entries, i.e. entries that will serve as the carriers of information. The second is the population of the entries with information of various kinds e.g. syntactic, semantic, phonological etc.

Our initial task, trivial as it may sound, is to concentrate on the first of these subtasks, creating what amounts to a word list, in a machine-readable and consistent format, for all the basic lexical entries of the language. The idea is that this will subsequently be used not only as a basis for applications (initially we will concentrate on spell-checking), but also as a tool for linguistic research on the language itself.

2 The Maltese Language

Maltese is the national language of Malta and, together with English, one of the two official languages of the Republic of Malta. Its use beyond the shores of the Maltese islands is limited to small emigrant communities in Canada and Australia, but within the geographical confines of Malta, the language is used for the widest possible range of types of interaction and communication, including education, journalism, broadcasting, administration, business and literary discourse.

Unsurprisingly in view of the disparate political and cultural influences the islands have been exposed to over the centuries, Maltese is a so-called 'mixed' language, with a substrate of Arabic, a considerable superstrate of Romance origin (especially Sicilian) and, to a much more limited extent, English. The Semitic (Western/Maghrebi Arabic) element is evident enough to justify considering the language a peripheral dialect of Arabic. Its script, codified as recently as the 1920s, utilises a modified Latin alphabet. This is just one of the peculiarities of

Maltese as compared to other dialectal varieties of Arabic, more important ones being its status as a 'high' variety and its use in literary, formal and official discourse, its lack of reference to any Qur'anic Arabic ideal, as well as its handling of extensive borrowings from non-Semitic sources. These features make Maltese a very interesting area for those working in the fields of language contact and Arabic dialectology.

2.1 The Maltese Alphabet

As noted above, Maltese is the only dialect of Arabic with a Latin script. Maltese orthography was standardised in the 1920s, utilising an alphabet largely identical with the Latin one, with the following additions/modifications:

Maltese	Pronunciation
ċ	chip (Eng)
ġ	jam (Eng)
h	silent
għ	mostly silent
ħ	hat (Eng)
ż	zip (Eng)
ie	ear (Eng) (approx)

2.2 Morphological Aspects of Maltese

The morphology is still based on a root-and-pattern system typical of Semitic languages. For example, from the triliteral root consonants *ħ – d – m* one can obtain forms like:

ħadem	work (verb);
ħaddiem	worker;
ħidma	work (noun);
nħadem	be worked (verb passive);
ħaddem	caused to work.

Most of these forms are based on productive templates (binyanim/forom/conjugations), of which Maltese has a subset of those in Classical Arabic. One other typical feature shared with Semitic languages is broken plural formation as opposed to so-called sound plural. A few examples are:

qamar	moon	*qmura*	moons;
tifel/tifla	boy/girl	*tfal*	children.

Plural formation in such instances involves a change in CV pattern. Sound plural formation involves affixation of suffixes such as -i, very common with words of Romance origin, -iet or -a as in:

karozza	car	*karozz – i*	cars;
ikla	meal	*ikl – iet*	meals;
ħaddiem	worker	*ħaddiema*	workers.

Maltese has taken on a very large number of Romance lexical items and incorporated them within the Semitic pattern. For example, *pizza*, a word of Romance origin, has the broken plural form *pizez* (compare Italian *pizza/pizze*), and *ċippa*, a very recent borrowing from English (computer chip) has a broken plural form *ċipep*. In certain cases, one gets free variation between the broken plural form and a sound plural based on (Romance) affixation, e.g.:

kaxxa	box	kaxex/kaxxi	boxes
tapit	carpet	twapet/tapiti	carpets.

The stem, as opposed to the consonantal root, also plays an important role in word formation, in particular in nominal inflection. Typical stem-based plural forms in which the stem remains intact are:

aħar	news item	*aħbar – iiet*	news
omm	mother	*omm – ijiet*	mothers

Verbs are also often borrowed and fully integrated into the Semitic verbal system and can take all of the inflective forms for person, number, gender, tense etc. that any other Maltese verbs of Semitic origin can take. For example:

spjega	explain (It. *spiegare*)
jispjega	he explains
nispjegaw	we explain
spjegat	she explained
spjegajt	I explained, etc.

ixxuttja	kick a football (Eng. *shoot*)
jixxuttja	he kicks
nixxuttjaw	we kick
ixxuttjat	she kicked
ixxuttjajt	I kicked, etc.

The vigour and productivity of these processes is attested to by the fact that one keeps coming across new loan verbs all the time (increasingly more from English), both in spoken and in written Maltese, without the language having any difficulty in integrating them seamlessly into its morphological setup.

Within the verbal system complex inflectional forms can also be built through multiple affixation. For example, the word *bgħat − t − hie − lu − x*

'I didn't send her to him', contains the the suffixes *−t* for 3rd person singular masculine subject (perfective), *−hie* for 3rd person singular feminine direct object, *−lu* for 3rd person singular masculine indirect object, and *−x* for verb negation. This ready potential for inflectional complexity is another Semitic feature of Maltese which applies across the board, whatever the origin of the verb. It also raises interesting questions concerning the nature of lexical entries, the relationship between lexical entries and surface strings, and the kind of morphological processing that is necessary to connect the two together.

Many of the linguistic issues that could help to resolve these questions are themselves unresolved for lack of suitably organised languaage resources (like the lexicon itself!). For this reason, we see the design/implementation of the lexicon, the development of language resources, and the evolution of linguistic theory for Maltese as three goals which must be pursued in parallel.

At this very early stage of the project, we have sidestepped many of the finer issues by opting to codify the most uncontentous parts of the lexicon first, as described below. At the same time, we are in the process of developing an extensible text archive which will serve as the basis for empirical work concerning both the lexicon and the underlying linguistics.

3 Constructing the Lexicon

The two main resources available to construct the lexicon are dictionaries and text corpora. Both, in some sense, are representative of the lexical behaviour of words, and both have their advantages and disadvantages.

3.1 The Dictionary Approach

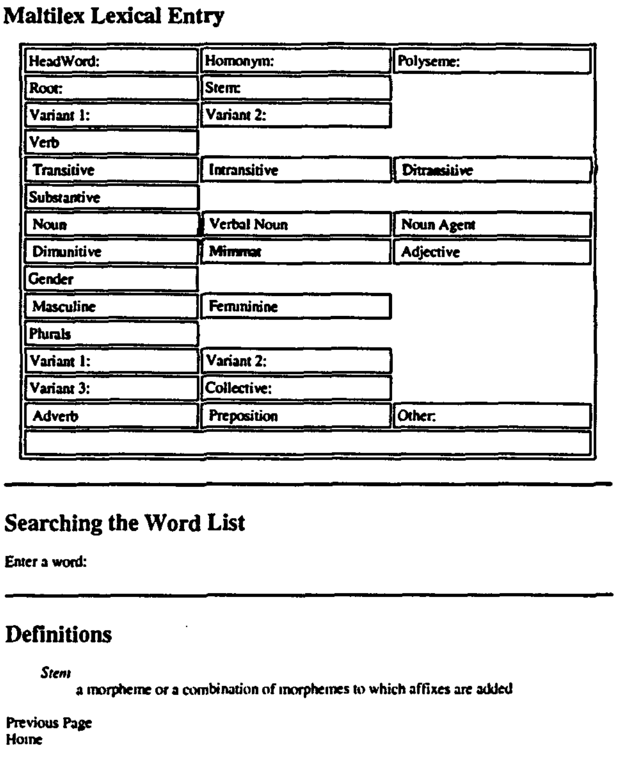
The basic idea underlying the dictionary approach is this: if some lexicographer has already gone to a great deal of trouble to compile a dictionary, why not make use of that work rather than repeat it? The appeal is obvious, and can be made to work, as is evidenced by, for example, the work of Boguraev and Briscoe (1987) who attempted to extract entries from the machine-readable version of Longman's dictionary. Problems of a practical nature soon arise, however, such as:

- What to do if a machine readable version of the printed dictionary is not available, as is in fact the case with Maltese.
- How to deal with the idiosynchratic formats adopted by different lexicographers, and how to handle the omissions and inconsistencies that are characteristic of all human oriented dictionaries.
- Once the information is available, how to represent it.
- How to deal with evolution of the language under investigation. Dictionaries always reflect the language as it was, not as it is. In the case of Maltese this problem is particularly acute, given that the most obviously useful dictionary contains a large number of entries that are regarded by many as archaic.

Many of these problems, except the last, are alleviated by adopting an essentially manual approach in the early stages. We have adopted the most complete and detailed dictionary currently available by J. Aquilina (Aquilina, 1987) and are in the process of transcribing the so-called *major* entries into our own format by means of a form interface as illustrated in figure 3.1. Major entries of this dictionary comprise a specific, orthographically distinguished (capitalised) subset containing the basic lexical forms of the language. They thus form a reasonable starting point for our purposes. The other (non-capitalised) entries are *derived* lexical forms of various kinds.

For the present, we are simply ignoring inflectional forms, since ultimately it is more efficient to assume that they can be systematically related to the basic entries by a morphological transformation of the sort implemented by Galea (1996).

The most important information is *headword*, a sequence of characters used to identify a particular lexical primitive or lexeme. Most of the time, the headword and the lexeme are in one:one correspondance, but there are exceptions. Distinct lexemes (and therefore entries) with the same headword are homonyms (e.g. *tikk*, a clock tick and *tikk*, a facial spasm). Single lexemes can also manifest polysemy, different meanings under the same headword (e.g.



Maltilex Lexical Entry

HeadWord:	Homonym:	Polyseme:
Root:	Stem:	
Variant 1:	Variant 2:	
Verb		
Transitive	Intransitive	Ditransitive
Substantive		
Noun	Verbal Noun	Noun Agent
Dimunitive	Mimmat	Adjective
Gender		
Masculine	Femininine	
Plurals		
Variant 1:	Variant 2:	
Variant 3:	Collective:	
Adverb	Preposition	Other:

Searching the Word List

Enter a word:

Definitions

Stem
a morpheme or a combination of morphemes to which affixes are added

Previous Page
Home

Figure 1: Internet Form for Dictionary Entries

tikka, a point-like diacritic mark and *tikka*, a very small amount).

These variations are accommodated using the headword (string), homonym (integer) and polyseme (integer) fields in the form, the integers deriving from the ordering implicit in the printed dictionary.

The second line of the form contains root (typically 3 consonants) and stem information for words of semitic and non-semitic origin respectively, whilst the third contains variants (e.g. *farfett*/*ferfett*, butterfly).

The remainder of the form contains mostly grammatical information, including that on (various forms of) plural. There is also space for comments from the individual lexicographer. The end product of the work described in this section is essentially a list of lexical entries for what we are calling the uncontentious parts of the language. The content of entries is essentially by reference (to the entries of Aquilina's dictionary) rather than literal.

3.2 The Corpus Approach

Comparatively recent technological changes have made it possible, in principle, to create and maintain corpora that are sufficiently large and accessible to be suitable for the purposes of lexical acquisition. One of the greatest advantages of the corpus approach to lexical acquisition, compared to the dictionary approach just described, is that in principle such corpora come as close as it is possible to get to a truly current snapshot of the language, particularly if they are continuously updated. Other arguments in favour of using texts as the basis for lexical acquisition are advanced in the editor's introduction to Boguraev and Pustejovsky (1995).

To adopt the corpus approach it is of course necessary to have a corpus, so that a priority task is the construction of a machine-readable, evolving record of the current written language. All the main Maltese language newspapers have been approached, and some journalistic texts (various fields) have already been obtained. We have recently managed to obtain speech corpora with parallel text of national radio news broadcasts. Furthermore, practical arrangements are currently being made for the provision of such materials on a regular and frequent basis. Book publishers have agreed to make titles from their respective ranges available for inclusion in the corpus. As it stands, the raw collection includes a number of book excerpts from various titles.

One feature of this approach is the constantly evolving relationship between corpus and lexicon: the corpus enriches the lexicon, but as the latter evolves, it can be used to add further information to the corpus in the form of annotations or tags, thus expanding its scope. A corpus annotated with part-of-speech tags, for example, can be used to infer a statistical model that can be harnessed to efficiently and automatically assign tags to previously unseen texts.

3.3 Character Representation

In the course of collecting corpus texts, it soon became apparent that, as a result of lack of standardisation early on in the introduction and spread of IT in Malta, a certain amount of anarchy reigns, with various computer/printer suppliers having developed and disseminated 'Maltese' adaptations of existing fontsets. The fact that they proceeded independently of each other and with no external regulation meant that the same Maltese-specific characters were assigned different ANSI codes in Windows (TTF) fonts supplied by competing sellers, making it difficult to read documents not only across platforms but also within the same platform.

A persistent challenge to the computational treatment of Maltese is therefore the question of text representation, i.e. the numerical coding for the characters that make up words. The requirements are:

- That the coding should follow an internationally recognised standard.
- That there exist appropriate fonts for use on the screen and on the printer across a variety of hardware platforms (PC/Mac/Unix).
- That there exists an accepted keyboard configuration to generate the codes.

Although no code satisfying all of these requirements exists, the most acceptable workaround available at present is to adopt fonts conforming to ISO8859-3, known as Latin Alphabet No. 3. Two PC-compatible fonts conforming to this standard are known as "Tornado" and "FTIMAL" and we are currently investigating the copyright status of each of these.

Given that these fonts are closely tied to PC (rather than Unix or Macintosh platforms), and given rather casual attitude taken to the adoption of text representation standards locally, we have defined a project-internal Standard Maltese Text Representation (SMTR) for storing text archives in a way that is (a) human-readable (and human-editable), (b) compatible with Unix systems and (c) easily translatable to and from any other coding format by means of simple finite-state methods (we are using Xerox's `xfst` for this purpose).

Maltese	Ascii
ċ	_c
ġ	_g
għ	_y
ħ	_h
ż	_z
ie	_i

4 Conclusion

This paper has attempted to convey our approach to the problem of rendering Maltese amenable to current language engineering techniques via the construction of a computational lexicon. One difficulty that we are currently facing is a shortage of appropriately qualified personnel to work on the project, though hopefully this problem will be alleviated by the appearance of our first CS/Computational Linguistics graduates during the coming year. Three subprojects are currently in the pipeline with the following themes:

- **Finite State Methods.** Development of finite state transducers for extracting lexical information from text corpora.
- **Computational Grammar.** Development of a grammar and parsing system for Maltese sentences. This will probably be based on HPSG.
- **Computational Morphology of Plural Forms.**

5 Acknowledgements

The authors grastefully acknowledge the contribution made by the Mid-Med Computer and Commerces Foundation to the funding of this project.

References

J. Aquilina. 1987. *Maltese-English Dictionary.* Midsea Books.

B. Boguraev and T. Briscoe. 1987. Large lexicons for natural language processing: exploring the grammar coding system of ldoce. *Computational Linguistics*, 13:203–218.

B. Boguraev and J Pustejovsky. 1995. *Corpus Processing for Lexical Acquisition.* MIT Press, Cambridge, Ma.

D. Galea. 1996. Morphological analysis of maltese verbs. Technical Report B.Sc Dissertation, Department of Computer Science, University of Malta.