# The Future of Maltilex

**Michael Rosner**

Dept. CSAI
University of Malta,
Msida MSD06, MALTA
mike.rosner@um.edu.mt

**Abstract**

The Maltilex project, supported by the University of Malta, has now been running for approximately 3 years. Its aim is to create a computational lexicon of Maltese to serve as the basic infrastructure for the development of a wide variety of language-enabled applications. The project is further described in Rosner et. al. (Rosner et-al 1999, Rosner et al., 1998). This paper discusses the background, achievements, and immediate future aims of the project. It concludes with a discussion of some themes to be pursued in the medium term.

## 1. Background

A pressing problem for Maltese is that of bringing it into the electronic age. This problem manifests itself in a number of different guises:

- Idiosynchratic conventions for the use of Maltese characters in documents. This includes character representations, sorting order, keyboard layout;
- Preference for using English in computer-based communications (e.g. email) and in interfaces. This is partly a question of habit, due to the physical awkwardness of using Maltese with computers. Another factor is the lack of technical vocabulary for most of the terminology associated with computer usage.
- Lack of formal linguistic knowledge as input for language-sensitive support, e.g. spell and style checking. There remains a serious lack of investment in the linguistics profession. Consequently, many areas of the language are not that well understood in comparison with better-studied languages.
- Lack of language resources. There is as yet no organised system for acquiring and cataloguing language resources. What has been acquired so far has been on an ad-hoc basis.

Concern with these issues have provided most of the impetus behind the Maltilex project.

### 1.1 The Language

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/MaghrebiArabic) 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 are 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.

The morphology is in part 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 (to work); ħaddiem (worker); ħidma (work/noun); ħadem (be worked/verb passive); ħaddem he caused to work.

Most of these forms are based on productive templates called *forom* of which Maltese has a subset of those in Classical Arabic. One other typical feature shared with Semitic languages is 'broken' plural. Plural formation in such instances involves an actual change in the pattern of vowels and consonants.

| **Singular** | **Plural** |
|---|---|
| qamar (moon) | qmura (moons) |
| tifel/tifla (boy/girl) | tfal (children) |

In contrast to this, sound plural formation involves affixation of suffixes such as -i (very common with words of Romance origin), -iet or -a as in:

| **Singular** | **Plural** |
|---|---|
| karozza (car) | karozz-i (cars) |
| ikla (meal) | ikl-iet (meals) |

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 (cf. 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.

| **Singular** | **Plural** |
|---|---|
| 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:

| **Singular** | **Plural** |
|---|---|
| aħbar (news item) | aħbar-ijiet (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, 'spjiega' (to explain) which clearly derives from the Italian 'spiegare':

| person | Singular | Plural |
|---|---|---|
| I | nispjiega | nispjegaw |
| II | tispiega | tispiegaw |
| III | jispiega | jispiegaw |
| III(fem) | spjegajt | - |

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 own morphological paradigms.

Within the verbal system complex inflectional forms can also be built through multiple affixation. For example, the word 'bgħatthielux' (I didn't send her to him), contains the the suffixes -t or 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 a host of 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 data - which could take the form of suitably organised language resources.

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.

## 1.2. Achievements

From a starting point characterised by an almost complete lack of computational and data resources, the project has achieved a number of aims including

### 1.2.1. Text Archive

The construction and exploitation of an electronic corpus of current Maltese has been an integral component of the Maltilex project since its inception. Text-representation posed one of the first problems we faced in building a corpus. In the absence of commonly agreed electronic standards for Maltese (the writing system extends the Latin alphabet to 29 letters) we adopted a purely ASCII system, whereby the five Maltese-specific characters (including one digraph) are represented by an `underscore + letter' combination.

The existing corpus (over 2,000 files, 1.5 million word-tokens) has grown in a relatively opportunistic manner, most of the data having been collected on the basis of what was accessible over the Internet. This has meant that written journalistic texts predominate: articles from one daily newspaper and three weeklies make up just under 90% of the global collection. The remainder consists of orthographically transcribed broadcast Maltese (radio and TV news, sport, and music programmes, including some telephone conversation between programme presenters and callers) and some literary works, both fiction and non-fiction.

This is by no means the final shape of the corpus. Our ultimate aim is that of building as representative a closed corpus as possible, the linguistic analysis of which would permit statements to be made about the current state of Maltese with a good degree of confidence. We are keenly aware of the need for greater balance in corpus construction.

The present collection represents an interim stage, designed to lay the groundwork not just of collecting, encoding and marking up a final corpus as such, but also of meeting project-wide immediate needs. One of the most pressing of these is the construction of an annotated machine-readable wordlist. In its present form, the corpus yields 57,840 word-types overall.

### 1.2.2. Tagsets

The design of a tagset involves principled decisions with respect to (a) the level of annotation required; (b) the level of granularity to be aimed for in the tagged corpus; (c) the categories of the object language – in this case Maltese - to be represented in the tagset; (d) the compatibility of the tagset to predetermined standards for the annotation of corpora.

Two tagsets for Maltese have been developed by J. Caruana and A. Gatt. Both are compliant with current standards for morphosyntactic annotation of corpora,

particularly the standards proposed by the Expert Advisory Group on Language Engineering Standards (EAGLES) and the Corpus Encoding Standard (CES). The development of two different tagsets for the same language facilitates further research into (a) the relationship between tagset structure/content and tagging accuracy; (b) the semi-automatic mapping of one tagset onto another.

Gatt's tagset is described in detail in Gatt (2001). Manual tagging of a sample corpus with this tagset is also underway. The latter procedure aims to provide a stochastic tagger with the necessary training data for subsequent automatic tagging of the full corpus.

### 1.2.3. LST

The output of a tokeniser is a flat list of tokens. One of the main problems confronting computational linguists faced with the output of a tokeniser on a large corpus is that of imposing structure on some tens of thousands of tokens. The question is: what structure is appropriate? There is no completely satisfactory answer to this question, because there are two kinds of structuring principle that we need to work with, one based on strictly defined morphological transformations which might be either Semitic or Romance, and the other based on similarity of meaning. The two principles do not always coincide. A trivial example of this phenomenon would be the words "kelma" (of Semitic origin meaning "word") and "kalma" (of Romance origin meaning "calm"). Clearly the two words belong to entirely different semantic fields, whilst at the same time having the same underlying consonant structure.

We have therefore decided to approach the problem of lexicon structuring from a more algorithmic standpoint as suggested in Micallef and Rosner (2000). The outcome is manifest in Dalli's Lexicon Structuring Technique, which identifies lemmas in an unstructured list of words without appeal to any predefined rules. It works using a set of statistical techniques adapted from bioinformatics algorithms that are usually used to structure genome data. To give a concrete example, the algorithm is capable of 'discovering' not just that the following words: kelma; kliem; kelmiet; tkellem; occurring within a much larger set, can be clustered together, but also what sequence of letters can be regarded as the optimal intersection that best characterises the set (it turns out to be "klm" which is very close to the Semitic root). Further details are to be found in Dalli (2002a).

### 1.2.4. Lexicon Server

The lexicon is not a storecupboard. For Maltilex, it is based on the view that the main point of a lexicon is to provide different kinds of *service* to different categories of human and non-human user.

For example, a common service provided by a lexicon is "lookup". A string is supplied and a definition of some kind is returned, if the word is present. Clearly, the way in which this operates will depend on the kind of user. An ordinary human user will probably require form-based input, with a carefully designed visual presentation of results. The interaction protocol would be quite different for non-human application programs. The lexical interface of a sentence parser, for example, should be completely functional, reliable and efficient. In many cases all the only result required is a list of possible categories. All this can be determined at the level of a suitably defined API. Besides lookup, a whole other range of other services is associated with the lexicon - maintenance and extension for example.

In short, the Maltilex view of the lexicon is as a collection of services delivered using different protocols. To accommodate this, we are currently experimenting with a new architecture. At the lowest is the core lexical information, stored in an efficient relational database. Basic lexicon services are delivered using a SOAP (Simple Object Access Protocol) server which provides XML-based interactions between different linguistic databases and systems over the HTTP protocol. Data records can be imported and exported in XML format and converted into efficient relational records transparently. WSDL (Web Services Descrption Language) is used to describe the services provided by the linguistic database system in a standard manner, significantly reducing the development time for the implementation of new clients. Finally, recent developments like UDDI may be used to facilitate the development of flexible and secure but easily accessible linguistic databases and processing resources. This architecture is further described in Dalli (2002b).

## 2. Future Developments

Maltilex has now reached a kind of watershed, in which some basic machinery, as described above, is now in place, but it remains largely unexploited. The objectives in the immediate future are therefore to build upon these foundations in order to pursue to broad strands of development: (a) to implement practical tools for document processing; (b) develop computational aspects of the linguistic theory of Maltese, and (c) further develop the lexicon itself to support (a) and (b). These are further discussed in the next sections.

### 2.1 Tools for Document Processing

The resources proposed for this project being somewhat limited, we have decided to focus on practical tools having a very close relationship with the evolving lexicon and which at the same time enjoy a certain level of coherence when considered together. The core application will be a spell checker that is flexible enough to operate with (a) different classes of document (eg scientific article versus business correspondence) and (b) different kinds of document source (e.g. keyboard, Optical Character Recogniser (OCR) and eventually, voice recognition. These are all associated with different classes of error.

### 2.1.1. Spell Checker

Spell checkers generally perform two distinct functions: (a) error detection, and (b) error correction. The first task may involve detection of non-words, for which different techniques are available, some of which have already been already investigated for Maltese. ome of these have already investigated at undergraduate level in the CSAI Department (Mizzi, 1999). Other errors may also be perfectly valid words. In such cases the fact that they are errors is defined by the context in which they occur. Sometimes this can be characterised using part of speech information (e.g "the answer is wring"). At other times more abstract semantic co-occurrence restrictions have been broken (e.g. Jupiter is the largest plant in the Solar System). Although is no single method that is guaranteed to provide an answer, experience suggests that effective error detection requires accumulation of evidence from different sources, followed by integration of the assorted evidence.

Once an error is detected, the next task is clearly to try and correct it. All correction methods have to generate and rank a list of candidate corrections, the most obvious way of doing this being to provide a simple list of common spelling errors paired with their corrections. Although such lists can be obtained by hand-correcting large quantities of text, such methods are extremely laborious. A less laborious alternative (which can work alongside the first method) is to employ a theory of error-generation and use it to generate the list of candidate corrections. For example, suppose we know that, while typing, letters that are physically adjacent on the keyboard are likely to be transposed. Then given "hte", we might be more inclined to generate "the" as a candidate than "het".

### 2.1.2. OCR System

We will also futher develop an Optical Character Recogniser (OCR) for Maltese which will also serve as another, badly needed, form of document input. This will build upon preliminary work already carried out (Felter, 2000) that was aimed at constructing a basic system which has the capability of giving good results (over 90% correct recognition) under ideal conditions. The main weaknesses of the current system are:

- **Recognition Errors**. The most frequent kinds of error are caused by confusion between certain pairs of visually similar characters: for example, "rn" is often confused with "m".

- **Isolated Character Recognition.** The existing system performs recognition by finding the physical location of the next character, and attempting to match it against stored character prototypes that are held in memory. There is no other source of information about what the next character might be. For instance, no attempt is made to relate the recognition of a single character to the surrounding context.

- **Stand-Alone System**. The current implementation has been developed as a stand-alone system. In order to increase its functionality it is important to fully exploit the possibilities of interaction with other tools and resources. For this to be possible a series of special application program interfaces API need to be developed.

Obviously, these shortcomings will be addressed in the proposed system. One particular theme that we will be pursuing is the capacity of the lexicon that we have been developing to serve as a source of information for improving the performance of the OCR/spell checker. There are basically two ways in which we intend to use it. First, it can, under certain circumstances, be used to guide the character recogniser or spell checker - specifically, when there is overwhelming evidence that a "known" word is in the process of being recognised. This guidance takes the form of hypothesised characters that the recogniser is expected to confirm or reject. As a general rule this task is far easier than recognising a character from scratch. The second way the lexicon can be used is as a basis for building a purely statistical model of word formation using HMM models.

## 2.2. Lexical Content

Lexical entries are currently structured under lemmas and are associated with a feature structure giving major category information and minor categories including person, number and gender. The core lexicon can be extended with other information that is considered relevant. In view of the proposal to be given below, extending the core lexicon with the following is desirable:

- Valency/argument structure information. e.g. *mar<intrans.>* vs. *ra<trans.>*;
- Quantitative information such as frequency of occurrence and productivity. Morphological word-formation processes are of particular importance, given the hybrid nature of Maltese morphology, with both Semitic and non-Semitic influences.
- Relative frequency of use of quasi-synonym pairs of Semitic vs. non-Semitic expressions. e.g. *għalliem/għalliema* vs. *teacher*/*sir/miss*
- Syntactic complement information. e.g. whether a verb takes a NP object or a PP object;
- Sortal restrictions arising from lexical semantic content. e.g. animacy and other information that restricts cooccurrence. e.g. ‘the stone ate the food’ is semantically marked since *stone* is semantically inanimate.

### 2.2 Morphosyntactic Description

One of the applications of the Maltilex lexicon is that of an aid to linguists in their morpho-syntactic analysis of language. The information present in individual lexical entries - particularly if extended along the lines suggested above - provides a rich repository of material that can be utilised in various ways by the linguist to create grammars that generate/recognise morpho-syntactic constructions.

In line with current methodology in formal semantics and syntactic analysis, we propose to focus on a 'fragment' of the grammar of Maltese with a view to providing an analysis within a computationally tractable linguistic framework. Specifically, we propose to develop a framework for the exhaustive study of the possessive noun phrase (NP) in Maltese.

The particular constructions that fall under the present proposal are the Construct State Construction and the Periphrastic Possessive Construction, the latter involving the preposition *ta'* ('of'). The exhaustive investigation of these constructions interfaces in crucial ways with research on formal and lexical semantics and the syntax of related constituents, such as determiners and quantifiers. Specifically, the following areas are of immediate interest:

1. the syntax of possessive constructions in Maltese and the differences between the two main types of construction outlined above;
2. the interaction of lexical semantic (sortal) and type-logical information in the determination of the syntactic structure of a possessive NP;
3. productive morphological and lexical processes related to nominals and their effect on the distributional properties of nouns in particular constructions;
4. the relationship between the possessive NP and other types of NP.

The outcome of this work will be a fully implemented HPSG grammar/lexicon.

### 3. Medium Term Future Developments

### 3.1. General Language Resources

Although we have succeeded in creating a sufficient basis for the developments reported above, the process of acquiring language resources for Maltese is still pitifully adhoc. We are therefore seeking to define the legal and scientific conditions under which a "Maltese National Corpus" could be set up as an ongoing initiative under which it would be possible to develop a national archive of tagged corpora for research and development, much along the lines of the British National Corpus.

### 3.2. Statistically Based Machine Translation

Malta is currently on the brink of a decision to join the EU. Whatever the political arguments for or against, it can be said with some certainty that this moment in time represents an unparalleled opportunity for statistical machine translation: currently, a huge volume of English/Maltese bitexts are either available or being produced. The Laws of Malta already exist in bilingual form; the *aquis communautaire,* a lengthy document describing the conditions of membership in the EU, is in the process of being translated by hand. In addition, a rather large quantity of English/Maltese bitexts are being produced by the local European Information Centre.

Our aim during the next two years is use bitexts as data for automatic acquisition of bilingual equivalences using alignment techniques such as those discussed in [8]. Later on we expect to make use of these equivalences in a series of bilingual translation aids.

## 4. Acknowledgements

This work is being supported by the University of Malta. Thanks also go to colleagues Ray Fabri, Joe Caruana, Albert Gatt, and Angelo Dalli all of whom are working actively for the project.

## 4. References

Dalli 2002a, Biologically Inspired Lexical Structuring Technique, Proceedings HLT Workshop, University of Pennnsylvania.

Dalli A. 2002b, Creation and Evaluation of Extensible Language Resources for Maltese, Proc LREC2002.

Felter, P. (2000), An Optical Character Recognition System for Maltese, CSAI Technical Report 2000-09, University of Malta

Gatt, A. (2001) Linguistic and Computational Aspects of the Design of an XCES-EAGLES, Compatible Tagset for Maltese. Technical Report, MaltiLex Project, University of Malta, 2001

Rosner, M. et al. (1999), Linguistic and Computational Aspects of Maltilex, Proceedings of the ATLAS Symposium, Tunis.

Rosner, M. et al. (1998) Maltilex, a Computational Lexicon for Maltese, Proc. Computational Approaches to Semitic Languages, COLING98, University of Montreal.

Melamed, I, (2001), Empirical Methods for Exploiting Parallel Texts, MIT Press 2001.

Micallef, P. and Rosner, M. (2000), The Development of Language Resources for Maltese, Workshop on Language Resources for Minority Languages, LREC2000.

Mizzi, R. (1999), The Development of a Statistical Spell Checker for Maltese, CSAI Technical Report 1999-18, University of Malta