Digital Language Resources and Tools for the Languages of Malta: A Roadmap

Consultation Document by the Committee for Information Technology, National Council for the Maltese Language

30th August, 2016

Albert Gatt (Chairman)
Mark Borg
John J. Camilleri
Ramon Casha
Michael Rosner

Executive Summary

This document details the work of the ICT Committee, set up by the *Kunsill Nazzjonali tal-Ilsien Malti* (National Council for the Maltese Language) during its 2013-16 tenure. It describes the work of the Committee on (a) a project involving the creation of an online dictionary of Maltese; (b) a proposal to collate, curate and enhance existing digital resources for Maltese.

Given the danger of *digital extinction* of the Maltese language, noted by Rosner and Joachimsen (2013; see Section 1 below), the present document makes the following proposals:

1. The creation of a central repository of language resources and tools related to Maltese, as well as the other languages used in the Maltese islands, notably English and Maltese Sign Language. The Committee itself has already undertaken steps to make such a repository available, as described in Section 2.1.
2. The setting up of an initiative, overseen by the National Council for the Maltese Language, to bring together stakeholders and ensure the long-term curation of language resources and tools in the Maltese context;
3. The involvement of more stakeholders and the sensitisation of the public as to the availability and importance of such resources.

1. Introduction

In a white paper published as part of a series under the auspices of the METANET4U European Initiative[1], Rosner and Joachimsen (2013) assessed the state of the Maltese language in the digital media, with particular reference to the extent to which it is used and, more importantly, the extent to which such use is supported through the deployment of tools and resources. Their conclusion was that the level of technological support for Maltese, compared to that for other European languages, is extremely low, giving rise to the threat of "digital extinction" for the language. The latter term refers to the possibility that Maltese, while possibly maintaining its status as a national tongue with institutional support and recognition, is nevertheless virtually absent from digital media.

This possible state of affairs can only be counteracted through a concerted effort to:

1. Enhance existing tools and resources, ensuring their continued availability, as well as updating them on a regular basis;
2. Develop new tools and resources for Maltese that are as yet unavailable, from software modules that support authoring (such as spell checkers) to more complex, "intelligent" systems that perform language-related tasks, such as machine translation, automatic summarisation and natural language generation;
3. Publicise such resources, ensuring that potential users are aware of their existence, while also minimising the effort needed on the part of such users to access and use them.

Since the ICT Committee was re-established as part of the National Council for the Maltese Language in 2013, it has made these issues the centerpiece of its programme of work for 2013–16.

The present document constitutes a report of the activities of the Committee during this period, together with proposals for the ongoing support of these goals. Specifically, it covers:

1. Documentation of the substantial number of resources that have been developed over the past several years by a variety of private and public bodies, bringing them together to provide a single entry point for interested users to identify the tools they need and assess their availability;
2. Launch of the Dizzjunarju Malti (DM) project, built from existing resources through a collaboration between various entities, and with the support of the ICT Committee of the National Council for the Maltese Language;
3. Suggestions for the kinds resources and tools which should be considered as the next priorities;

[1] See http://www.meta-net.eu/whitepapers/volumes/maltese

4. Proposal of a central repository that will support the curation of tools and resources over the long term, in order to ensure continued availability, facilitate updating, and support the creation of new resources;
5. Identification of potential stakeholders who would benefit from the creation of such a support structure.

Each of these items is covered in the following sections.

2. Existing resources

Over the past several years, a large number of resources and tools for the languages of Malta[2] have been created, varying in size and breadth of scope. These tools and resources have been created through initiatives within the research community, especially at the University of Malta, as well as by private individuals and the private sector, sometimes in collaboration with national bodies.

Broadly speaking, these resources and tools can be classified as follows:

1. **Lexicons:** To date, a variety of lexical resources have been developed, ranging from small-scale wordlists to large-scale collections of lexical entries with associated morphological information. In addition, a lexicon of Maltese Sign Language is currently in progress, as part of a research initiative spearheaded by the Institute of Linguistics, University of Malta.
2. **Corpora**: Corpora of the Maltese language have been in existence at least since the public launch of the Maltese Language Resource Server in 2009. Corpora now include a large corpus of written Maltese in a variety of genres, as well as a corpus of Learner English in Malta. In addition, other text collections have also been created for specific purposes. These corpora have different modes of access (including, for example, publicly searchable web interfaces) as well as different levels of annotation (including, for example, part of speech tagging).
3. **Software tools**: There is also a significant number of software resources that handle Maltese, ranging from text annotation tools such as tokenisers and part-of-speech taggers; authorship support tools such as spell checkers; and end-to-end applications such as the text-to-speech synthesis system produced by CrimsonWing.

2.1 Collection of current resources with a single access point

As one if its first tasks, the Committee took it upon itself to compile **a list of all known resources**, collected from the personal expertise of its members. It is the view of the Committee that these resources should be made accessible from one single access point which is public and available online. While such an access point need not imply that all resources are hosted on the same server, given that some individuals and/or organisations may wish to maintain some level of control over their distribution, a portal which is continuously updated would

[2] The languages of Malta include not only Maltese, but also Maltese English – the specific variety of English that has evolved on the Maltese islands – and Maltese Sign Language.

mean that interested members of the public and potential users have a one-stop shop to consult and identify the right tools given their needs.

Such an access point has been set up as part of the existing Maltese Language Resource Server[3], hosted by the University of Malta and maintained by the Institute of Linguistics. It lists all known resources and links to them, where public URLs are available.

2.2 Dizzjunarju Malti (DM)

In addition to the compilation of digital language resources and their being made available, the Committee has spearheaded the *Dizzjunarju Malti* (DM)[4], which is a collaboration among the following entities:

- The ICT Committee of the National Council for the Maltese Language;
- The Institute of Linguistics at the University of Malta, where most of the content development took place;
- The Malta Communications Authority, whose primarily role was, and continues to be, the financial support and overall management of the project;
- The Vodafone Foundation, which provided funding for the project;
- Infusion Ltd, which was responsible for the development of the public interface and accompanying mobile apps, while also providing some support in kind.

To date, the project has had funding amounting to approximately €10,000, although efforts are underway to secure further funding for ongoing update and maintenance work. The money received to date was spent on hiring research assistants to correct and update content on the existing Ġabra lexicon, as well as pay for the development of a user-friendly public interface.

The public press launch of DM was made on March 29, 2016, at the Malta Communications Authority headquarters, in the presence of MCA officials, Prof. Juanito Camilleri, Rector of the University of Malta, and the chairman and secretary of the National Council for the Maltese Language.

At the time of writing, the DM dictionary contains some 15,200 entries. The website receives some 1000 hits per day. Suggestions for new entries are submitted daily by users.

2.3 Next priorities

Given the current resources and tools available for Maltese (listed on the portal referred to in Section 2.1 above), the Committee would like to give the following list of suggestions for important technologies which are still lacking for Maltese:

1. Spell checker with wide coverage of inflected and suffixed forms;
2. Machine translation system to/from English and other languages;
3. Speech recognition;

[3] http://mlrs.research.um.edu.mt/resourcefinder/
[4] http://www.maltesedictionary.org.mt/

4. Parsing to phrase-structure or dependency representations;
5. Treebanks — corpora annotated with tree structures, often used for machine learning tasks;
6. WordNet[5] — lexical resource where entries are grouped into cognitive synsets, with detailed lexical relationships between them.

A number of the above tools and resources are active topics of research at the University of Malta, notably within the Institute of Linguistics and the Department of Intelligent Computer Systems. Further work has also been carried out by private individuals and/or companies.

Nevertheless, such efforts remain disparate. Furthermore, prior to the launch of the portal listing all language resources as a result of this Committee's initiatives, there has been no single access point where all such resources could be identified.

3. The way forward

The need for more coordinated efforts on the creation and maintenance of language resources in the Maltese context is increasingly evident. Frequently, efforts appear somewhat piecemeal, or remain at a remove from public consciousness. This is a result of two principal factors:

1. Projects in industry and academia, which produce such resources as primary or secondary deliverables, are usually time-bound and dependent on limited funds. As a result, once projects reach their termination, maintenance and publicity grind to a halt.
2. Potential industrial partners, public bodies, and private individuals are under-informed about the value and utility of language resources and the added value they bring to commercial and research enterprises.

These observations motivate the proposals and recommendations made immediately below.

3.1 A National Digital Repository for the Languages of Malta (NDRLM)

Beyond the short-term desideratum of a single access point for interested users and stakeholders, there is a further need that will need to be addressed in the medium term, namely, the creation of **a structure that will provide ongoing logistical and technical support for the maintenance and update of these resources.** Such a structure should result in the creation of a national repository. Concretely, this can be envisaged as a portal through which different tools and resources can be accessed, and on which new resources and tools can be announced and hosted. The existence of such a national repository would also facilitate the identification of new stakeholders and enhance co-ownership of resources.

[5] See https://wordnet.princeton.edu/

3.2 Core principles

The Committee believes that in order for this repository to achieve its ultimate aims, it needs to be based on the following principles:

- **Promote openness**: Resources should be free for use *for all purposes*. This means that elements of the repository should be licensed accordingly. Note that this means, for example, that users may avail themselves of such resources for research *or* commercial purposes. In both cases, we believe that lowering the hurdles for use as much as possible would increase the chances of (a) development of novel tools and resources that rely and build upon those already in existence; and (b) improvement on existing resources.
- **Common Formats**: While data formats will differ depending on the nature of a specific resource, a common format for the representation of metadata is crucial to enable services such as searchability and cataloguing. In developing such a common format, it is possible to rely on numerous precedents, notably large-scale endeavours such as METANET4U and CLARIN (to which the national repository may ultimately contribute).
- **Searchability:** A portal should provide search facilities which are easy to use for users of different levels of expertise.
- **Seamless integration:** Developers who wish to integrate or exploit existing tools within new systems and architectures should be able to do so easily. This can be achieved through the development of common interfaces to tools hosted as part of the repository, especially web services. This is a desideratum and could be interpreted in terms of a set of guidelines, rather than imposed, since not all tools/resources may be able to conform to the same interface standards.
- **Preservation**: efforts should be made to ensure software preservation and maintenance (e.g. ensuring compatibility with new versions of operating systems), as well as accurate documentation.

3.3 Support structure

If this proposal is to come to fruition, it will need the support of a number of significant players, notably:

- The National Council for the Maltese Language.
- The University of Malta, that has played a key role in the development of language resources and tools, through the efforts of various departments and institutes.
- Industrial players, consisting of both companies that have already evinced an interest in the field and have contributed to its development, as well as companies with a potential interest in such developments.
- The Government of Malta, through the action of its constituted bodies, especially agencies such as the Malta Communications Authority (MCA) and the Malta Information Technology Agency (MITA).

In practice, the Committee believes that these desiderata can be achieved through the establishment of a partnership between the University of Malta and public and private bodies. A way forward for such a partnership is already prefigured by the type of collaboration set up between the Malta Communications Authority and the University in its DM project (see Section 2.2 above).

Specifically, the Committee proposes that the National Council for the Maltese Language undertake responsibility for:

1. Providing long-term funding for the maintenance and continuous update of the existing Maltese Language Resource Server, which should henceforward be viewed as a national, rather than an exclusively university-based (that is, academic) repository;
2. Setting up a body made up of members of the academic and industrial partners, as well as Government agencies, to manage and oversee the update of these resources.

3.4 Potential stakeholders

If a national framework for the curation of language resources is available, it should serve to facilitate the involvement and co-ownership of new stakeholders. Such potentially interested parties, who would be both beneficiaries and contributors, include operators who may occupy one or more of the following roles: *producers, consumers* and *maintainers of resources.*:

- **Producers** include academic or industrial partners or private individuals who are engaged in developing new resources and tools;
- **Consumers** are any private individuals, public or commercial bodies who make use of resources and tools in their day-to-day activity;
- **Maintainers** include any individuals or organisations who actively contribute to the further development of existing resources and tools.

The following is a tentative and partial list of existing organisations which, in all likelihood, have an active interest in being involved in one or more of these roles:

- The press, including Maltese and English-language newspapers, book publishers, television and radio stations, as well as operators in the new media;
- Organisations involved in archiving, such as the National Archives, who have an interest in the digitisation of resources (many of which are ultimately linguistic in nature) as well as the Curia (which is beginning to digitise its archives), the Court Services, Housing Registry etc;
- Organisations whose day-to-day activities involve the use of language in specialized domains and who would therefore benefit from the existence of such resources. These include, but are not limited to, firms and organisations engaged in legal activities such as contract drafting and translation;
- Translators and interpreters, including the members of the Maltese translation bureaus within the institutions of the European Union;

- Educators and learners of Maltese;
- Researchers in language and linguistics and language technology;
- Software developers who want to provide user interfaces and/or any kind of natural language processing features involving Maltese.

4. Conclusions

It is the belief of this Committee that the time is ripe for concrete initiatives to be undertaken to consolidate and further develop digital resources for the languages of Malta. This can only happen effectively with a concrete support structure of the kind described in Section 3.3, which will be a first step towards addressing the needs identified in Section 3.1 and in widening the circle of stakeholders who are involved in the curation of the next generation of tools and resources for the languages of Malta.

5. References

Rosner, M. and J. Joachimsen (2013). The Maltese Language in the Digital Age / Il-Lingwa Maltija fl-Era Diġitali. (META-NET White Paper Series). Berlin: Springer. Available from: http://www.meta-net.eu/whitepapers/volumes/maltese

The Maltese Language Resource Server. Available at: http://mlrs.research.um.edu.mt