Digital Corpora and Other Electronic Resources for Maltese

Albert Gatt
Institute of Linguistics
University of Malta
albert.gatt@um.edu.mt

Slavomír Čéplö
ÚČNK
Charles University
bulbul@bulbul.sk

1 Introduction

This paper describes the development and current status of digital corpora and other NLP resources for Maltese. After an introduction which briefly examines the linguistic situation and the history of corpus development for Maltese, the paper will focus on the efforts of two teams of researchers to create a digital corpus and set of related tools for contemporary and historical Maltese. These efforts, while largely independent at the outset, are in the process of being aligned.

2 Corpus Linguistics and Maltese

The groundwork for current initiatives to collect machine readable data for Maltese was laid in the MaltiLex Project (Rosner et al., 2000). The stated aim of Maltilex was to construct an electronic Maltese lexicon, based on corpus data; however, the resulting corpus was of a relatively small size and lacked any meaningful structural or grammatical annotation. More recently, Ussishkin et al. (2009) used web resources to create a medium-sized corpus, primarily for use in the extraction of lexical resources to inform experimental work on Maltese lexical processing. Two recent European initiatives, Clarin and METANET4U have also provided impetus for further development of Maltese language resources: while work within Clarin was mainly focused on the digitisation of resources within the humanities, the METANET initiative aimed to build a common, Europe-wide infrastructure to accomodate corpora and text and speech processing tools.[1]

The present paper focuses on two more recent efforts to build corpora and related tools for Maltese, on a much larger scale. We summarize the challenges involved in data collection and text preparation as well as the choice of corpus management tool and related issues. We focus especially on issues related to opportunistic data collection from the web, and a description of the methods used to harvest text from web pages and online documents in various formats. The resulting corpora – the *MLRS Corpus*[2] running on the IMS Open Corpus Workbench, recently released in version 2.0 with part of speech tagging, and the beta version of the *bulbulistan corpus*[3] based on the NoSketchEngine – are then briefly introduced with some comments on the impact of *MLRS Corpus* as a linguistic resource on both researchers and the lay public. While *MLRS* and *bulbulistan* arose as separate initiatives, efforts are underway to make them compatible.

3 Data Preparation and Annotation

The next section goes on to describe the process of adaptation of the data in some detail. In particular, we describe

(i) the creation of a spell-checking dictionary through crowd-sourcing and its use in the creation of version 2.0 of the MLRS corpus;
(ii) the use of a dictionary-based algorithm to correct Maltese text written without diacritics (a common practice whereby, for example, *ħ* is written as *h*) and alternative ways of dealing with this problem;
(iii) ongoing efforts to harness these resources to develop better spell-checking algorithms;
(iv) the addition of new levels of annotation in the corpora.

Annotation is then discussed in some detail with special attention devoted to POS-tagging and related issues. These include the choice of tagset with regard to Maltese as a Semitic language with a hybrid morphology (including a highly productive, non-Semitic component), the choice of tagger and the process of POS-tagging with minimal available manually tagged data. The resulting versions of both corpora (the current, tagged version 2.0 of the *MLRS Corpus* and the newly released tagged beta version *bulbulistan corpus*) are then discussed. We also consider the benefits of a multi-level approach to POS annotation, whereby text is first tagged with basic, category-level information, with subsequent morphological analysis to include more fine-grained POS-level information (such as number, gender, and pronominal suffixes).

4 Balance and Representativeness

One of the challenges with both *MLRS Corpus* and the *bulbulistan corpus* arises from their being opportunistic corpora. This is chiefly manifested in

[1] Several tools developed for the Maltese Language Resource Server, one of the corpora discussed here, are now available as web services under the METANET4U framework, including a POS Tagger, tokeniser and sentence splitter. See http://metanet4u.research.um.edu.mt

[2] Maltese Language Resource Server, Malta, accessible at http://mlrs.research.um.edu.mt/.

[3] Bratislava / Prague, accessible at http://www.bulbul.sk/bonito2/.

their composition where journalistic texts are overrepresented and some text types and genres are represented scarcely or not at all (cf. Table 1 and 2).

Text type	Number of tokens
Journalistic texts	68.800.000
Parliamentary debates	43.400.000
Belles lettres	375.000
Academic texts	170.000
Legal texts	4.800.000
Religious texts	403.700
Speeches	18.000
Web pages (blogs etc., including Maltese Wikipedia articles)	6.500.000
Miscellaneous other texts	123.000

Table 1: MLRS corpus

Text type	Number of tokens
Journalistic texts	80.000.000
Parliamentary debates	50.000.000
Belles lettres	400.000
Academic texts	100.000
Other (blogs, ads etc.)	50.000

Table 2: bulbulistan corpus

We discuss current efforts and future plans for transition from this model to a more balanced and representative one with some attention devoted to the bilingual nature of Maltese society and what this means for the ideas of representativeness and balance (considering e.g. the comparatively low proportion of texts in some subject areas such as economics or mathematics as compared to other languages of similar size and status). Various proposals for creating a balanced and representative subcorpus are introduced and the methods for their creation are proposed.

5 Diachronic Dimension

In their current versions, both corpora are primarily synchronic, but efforts have been made to add a diachronic dimension by including older texts. We will discuss these efforts and the theoretical and practical challenges they pose. These involve diverse issues ranging from the definition of what counts as older (currently the focus is on the period between 19th century and the establishment of official orthography in 1921) through selection and collection of data to its annotation. This includes dealing with various systems of orthography which for the most part exhibit significant variation and inconsistency and in some cases use non-Latin characters, all of which raises interesting problems in text normalization and processing.

6 Beyond Corpora

In the course of compiling the corpora, a number of related tools and resources have been created to aid in the computer-assisted processing of Maltese. The tools comprise a sentence splitter, chunker, tokenizer and POS-taggers, while the resources include full list of Maltese verbs adapted to the Semitic root structure (based on Spagnol 2011) and a library of out-of-copyright literary works in Maltese, all publicly available.

We also discuss existing and forthcoming work on solutions for computer-aided morphological and syntactic analysis, such as the inclusion of Maltese in the Grammatical Framework project (Dannélls and Camilleri 2010).

7 What's Next

In lieu of a conclusion, we lay out a brief road map for further development of both corpora, including the gradual addition of more and more diverse texts and further levels of linguistic description (lemmatization, morphological analysis and rudimentary syntactic description) as well as the prospects of merging them into a single resource. We also discuss the development of other electronic resources for Maltese and the inclusion of Maltese corpus data in other projects, such as the InterCorp project and the SketchEngine resource.

References

Bovingdon, R. and Dalli, A. 2006. "Statistical analysis of the source origin of Maltese." In: A. Wilson, D. Archer and P. Rayson (eds.) *Corpus linguistics around the world*. Amsterdam: Rodopi.

Dannélls, D. and Camilleri, J.J. 2010. "Verb Morphology of Hebrew and Maltese - Towards an Open Source Type Theoretical Resource Grammar in GF." In: *Proceedings of Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects (LREC 2010)*. Malta, April 2010.

Rosner, M., Fabri, R. and Caruana, J. 2000. *Maltilex: A Computational Lexicon for Maltese*. Msida: University of Malta.

Spagnol, M. 2011. *A Tale of Two Morphologies: Verb structure and argument alteration in Maltese*. Unpublished PhD thesis, University of Konstanz. Available online at http://d-nb.info/1017360529/34.

Ussishkin, A., Francom, J. and Woudstra, D. 2009. "Creating a web-based lexical corpus and information-extraction tools for the Semitic language Maltese." In: *Proceedings of the SEPLN-SALTMIL 2009 Workshop: Information Retrieval and Information Extraction for Less Resourced Languages*, University of the Basque Country, 9-16.