Chapter 27
Language Report Maltese

Michael Rosner and Claudia Borg

Abstract This chapter is a highly abbreviated version of an update (Rosner and C. Borg 2022) to the META-NET White Paper on Maltese (Rosner and Joachimsen 2012). Like its predecessor, the update forms part of a series for all European Languages. Section 1 provides a brief description of the language, its national status, its general typology as a language, and its current usage in the digital sphere. Section 2 gives an overview of technologies and resources that are currently available. Finally, Section 3 frames the main shortcomings of Maltese language technology in terms of fragmentation, and offers some recommendations on how that might be reduced.

1 The Maltese Language

Maltese (il-Malti) is an official EU language and the national language of the Maltese archipelago. 97% of the Maltese population (ca. 400,000 people) consider it their mother tongue. It is also spoken by communities in Australia, Canada, the USA and the UK. Maltese is derived from late medieval Sicilian Arabic with Romance superstrata, and is often referred to as a mixed language due to the large number of loan words from Italian, English and French. It shares characteristics with other Semitic languages, making use of root-and-template morphology whereby various forms of the same lexeme are formed by interdigitating vowels between a fixed sequence of root consonants. The main distinguishing characteristics of Maltese are free word order, mixed morphology, aspect-based temporal system, and lack of a morphological infinitive. Unlike other Semitic languages, the Maltese alphabet is based on the Latin one with the addition of some letters with diacritic marks and digraphs (ċ, ġħ, ż, ġ, ħ). It contains 24 consonants and 6 vowels. According to Fabri (2011), the writing systems used for Maltese were somewhat ad hoc before 1920, but a degree of consistency among writers and in publications became a reality in the 1950s.

Within the digital sphere, there have always been several Maltese language newspapers. The broadcast media (radio and TV) are almost exclusively in Maltese. Since

Michael Rosner · Claudia Borg
University of Malta, Malta, mike.rosner@um.edu.mt, claudia.borg@um.edu.mt

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_27

the previous report, there has been a general decline in hard-copy newspaper readership, as all the media are now available online and the majority of readers prefer the online version. Various online-only news websites have appeared, one of which (Newsbook) operates bilingually. The full Maltese character set is now universally used. Social media are extremely popular (97% of the population according to a 2021 survey). Facebook remains the most accessed, but there is a trend of increased usage of Instagram and YouTube. Unlike other EU countries, Twitter usage in Malta is remarkably low. The Maltese Wikipedia currently ranks at 204/325 (for comparison, English, Portuguese, Irish, Icelandic, Romansch rank at 1, 18, 93, 95, and 213, respectively). It contains nearly 4 million words distributed over 4,400 content pages (cf. 6.5 million for English). This compares to about 3,000 pages in 2011; there are ca. 19,000 registered users with only about 40 active users (making changes every 30 days or less. YouTube gives rise to localised content in many other countries but the local website still operates predominantly in English. In general, there tends to be a gap between social media content creators and non-creators. However, a renowned online page which has successfully bucked this trend is Kelma Kelma which started in 2013 as a Facebook page and gathers many interesting original contributions by locals about the Maltese language. The top-level country domain for Malta, .mt, is administered by the Malta Internet Foundation, has currently ca. 17,000 domain names and subdomains, more than three times the figure in 2010.

2 Technologies and Resources for Maltese

Rosner and Joachimsen (2012) describe the main enablers and contributions to Maltese Language Technology up to ca. 2011. 2012 marked the public release of the MSE speech synthesiser (M. Borg et al. 2014), whilst Gatt and colleagues began revamping the University's MLRS resource server (Rosner 2008; Gatt and Čéplö 2013) to include semi-automated data-collection, a tagger, Korpus Malti v3.0 (2016), containing ca. 250 million annotated tokens, pattern-based search facilities, CLEM, a 1 million token Corpus of Learner English in Malta, Ġabra, an Open Lexicon for Maltese, and a Dictionary of Maltese Sign Language.

Most available corpora are monolingual written text. A few are spoken, and fewer still are multimodal such as MAMCO (Paggio et al. 2018). Many monolingual corpora form part of unannotated *multilingual* collections. Others are by-products of projects and annotated for MWE identification (PARSEME) or POS Tagging (MLRS), anonymisation (MAPA), morphological analysis (UniMorph), NER (WikiAnn) etc. Bilingual/multilingual resources include the Laws of Malta, the Government Gazette, and the Acquis Communautaire.

Regarding tools and services, besides low-level text preprocessing for tokenisation, sentence and paragraph splitting and POS-tagging, the Ġabra dictionary has evolved into the online Dizzjunarju tal-Malti app. Machine translation for Maltese has improved not only through the availability of free tools like Google Translate, but also as a result of DGT's eTranslation platform whose increased takeup by pub-

lic administration officials followed a series of workshops organised through ELRC. Much recent effort has been focused on dependency parsing and ASR. There is now a 2000-sentence Universal Dependency Treebank for Maltese which has supported experiments (Zammit et al. 2019) aimed at delivering a prototype dependency parser in 2022. Similarly, for speech technology, the locally funded MASRI project has delivered a fully annotated speech corpus (Hernandez Mena et al. 2020). Most resources mentioned above are freely available through MLRS and also EU platforms.

Currently, the main drivers for the evolution of future Maltese LT are targeted national initiatives, against a mixed background of projects at EU level. At the national level, the National AI Strategy (2019) focuses on the creation of an AI ecosystem infrastructure including tools to enable Maltese Language AI solutions, with funds earmarked for Maltese LT resources. The Malta Digital Innovation Authority (MDIA) is committed to supporting Maltese LT tools which will focus on morphological analysis, dependency parsing, named entity recognition and POS tagging. In 2019, the Government also committed funds to the development of a spell checker. However, there is no information with respect to the progress of this important initiative. Meanwhile at the EU level Maltese participation in a wide range of projects, actions and initiatives including ELE, ELG, ELRC, DARIAH, LCT, LT-Bridge, MAPA, Nexus Linguarum, and NLTP, has ensured a level of Maltese presence on the European scene and also produced some specialised resources and tools.

3 Recommendations and Next Steps

Maltese LT is indeed alive, but manifests an important weakness: it is highly fragmented, in different ways: 1. between national efforts (small-scale, Maltese-focused) and international ones (large-scale, language-independent); 2. across resources/tools which are not necessarily compatible with each other; and 3. between users and developers of LTs (reduces the perceived relevance of the technologies developed). To address these requires further investigation of techniques like transfer learning, as seen, for example, in the MAPA project where general language models were successfully used for Maltese NER. Issue 2. can be reduced by insisting that such resources inhabit a framework which includes the necessary protocols to ensure interoperability, as seen in European infrastructures like ELG and NLTP, funded under CEF, aiming to build a National Language Platform for Maltese integrating eTranslation services developed by the European Parliament with fine-tuned local translation memories, and providing a central point for collecting different LT services together. 3. is in part the result of insufficient involvement of the IT industry in LT. Despite the latter being a major component of the local economy, the number of technical LT providers is very low. LT has a crucial role to play as a natural bridge linking IT, AI, communication and multilinguality. More needs to be done to support that role by encouraging participation in ELG by local IT players, among others. In 2016, the IT subcommittee of the Council for the Maltese Language had recognised the need for the long-term curation of resources, recommending the creation of a central

repository, and efforts to involve more stakeholders concerning the availability and importance of resources. Some progress towards the realisation of these recommendations has been made but the effort needs a substantial and sustained coordinated investment across the different sectors involved.

References

Borg, Mark, Keith Bugeja, Colin Vella, Gordon Mangion, and Carmel Gafa (2014). "Preparation of a Free-Running Text Corpus for Maltese Concatenative Speech Synthesis". In: *Perspectives on Maltese Linguistics, Studia Typologica 14*. Ed. by Albert Borg, Sandro Caruana, and Alexandra Vella, pp. 297–318.

Fabri, Ray (2011). "Maltese". In: *The Languages of the 25. Revue belge de Philologie et d'Histoire: RBPH*. Ed. by Christian Delcourt and Piet van Sterkenburg. Amsterdam, Philadelphia: John Benjamins, pp. 17–28.

Gatt, Albert and Slavomír Čéplö (2013). "Digital corpora and other electronic resources for Maltese". In: *Proceedings of Corpus Linguistics*. Ed. by Andrew Hardie and Robbie Love. University of Lancaster, UCREL.

Hernandez Mena, Carlos Daniel, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani (2020). "MASRI-HEADSET: A Maltese Corpus for Speech Recognition". In: *Proceedings of LREC 2020*. Marseille, France: ELRA, pp. 6381–6388.

Paggio, Patrizia, Luke Galea, and Alexandra Vella (2018). *Prosodic and gestural marking of complement fronting in Maltese*. DOI: 10.5281/zenodo.1181805.

Rosner, Mike (2008). "Electronic Language Resources for Maltese". In: *Proceedings of Bremen Workshop on Maltese Linguistics*. Springer.

Rosner, Mike and Claudia Borg (2022). *Deliverable D1.25 Report on the Maltese Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-maltese.pdf.

Rosner, Mike and Jan Joachimsen (2012). *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/maltese.

Zammit, Andrei, Slavomír Čéplö, Lonneke van der Plas, and Claudia Borg (2019). *A Dependency Parser for Maltese: Comparing the impact of transfer learning from Romance and Semitic Languages*.

Chapter 28
Language Report Norwegian

Kristine Eide, Andre Kåsen, and Ingerid Løyning Dale

Abstract The use of Language Technology (LT) has greatly increased in Norway in recent years, as have the linguistic resources needed to make them work. In the past 10 years, Norwegian has adopted new or improved versions of machine translation, speech technology, chatbots and digital assistants, and machine learning has improved. Nevertheless, LT for both written standards of the Norwegian language – the majority Bokmål and minority Nynorsk – is nowhere near the same level as that of major European languages such as English, German, French and Spanish.

1 The Norwegian Language

Norwegian is a North Germanic, verb second, SVO language, spoken by about five million people in Norway, with some additional speakers in the Norwegian diaspora in the US and South America. Norway is a highly digitalised society.

There is great dialectal variation in Norway, and dialects have much higher prestige than in the other Scandinavian countries. Unlike other official European languages, there is no official standard for spoken Norwegian. People tend to speak their own dialect, and expect to be understood. This dialectal variation as well as the pitch accent found in most dialects present the biggest challenges for Norwegian speech technology. While there is no official standard for the spoken language, there are two official written Norwegian languages, Bokmål and Nynorsk. The minority language, Nynorsk, has about 500,000 speakers. All public bodies at state level must be able to correspond with citizens in both written standards, and even though the linguistic differences between Bokmål and Nynorsk are rather small, most types of language technology, such as machine translation, chatbots, spellcheckers, speech-to-text and text-to-speech, need separate tools for each language. Both standards reflect dialectal variation and allow for large formal morphological as well as orthographic variation.

Kristine Eide
The Language Council of Norway, Norway, kristine.eide@sprakradet.no

Andre Kåsen · Ingerid Løyning Dale
The National Library of Norway, Norway, andre.kasen@nb.no, ingerid.dale@nb.no

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_28

With this variation, in combination with highly productive compounding, one single word can have a relatively high number of different spellings, which is a challenge for language technology (Smedt et al. 2012a,b).

2 Technologies and Resources for Norwegian

The overall accessibility of Language Resources (LRs) for Bokmål is fairly good (Eide et al. 2022). Size and contemporaneity are in place for unstructured and semi-structured data. With good linguistic insight, one can build several specialised applications and services from openly available resources. In contrast, most types of LRs and LTs are either scarce or lacking for Nynorsk, although both speech and text data have been added to the largest, open repository for language data (Språkbanken) in recent years. Domain-specific data is severely limited for both Bokmål and Nynorsk. This is also true for the spoken language with all its dialectal variation.

Awareness of the differences between Nynorsk and Bokmål is low outside Norway's borders. Norwegian can often be found in large, multilingual LR collections, and is available as a language choice also on large online platforms. However, both nationally and internationally developed tools and services cater first and foremost to the Bokmål written standard, or the Eastern Norwegian spoken dialect.

Speech technology development is challenged by the dialectal variation, in addition to the two orthographic standards that often allow for spelling variations. There are pronunciation lexicons which cover Bokmål and Nynorsk orthographic forms, and dialectal variation in pronunciation transcriptions is under development for both. Some speech corpora with dialectal variation and a mix of read and spontaneous speech exist, some have transcriptions in both standards. These corpora have proven useful in improving speech recognition scores, but they are either not large enough, or somewhat lacking in domain, style, societal or situational variation to train a robust general purpose speech recognition system. Until recently, speech processing tools have been almost non-existent for Nynorsk. Those that are deemed usable, for either of the written standards, are in general proprietary and not freely available.

The largest text corpus is the Norwegian Colossal Corpus (NCC), which comprises a majority of all Norwegian published works (digitised using OCR), in addition to several other corpora, including Wikipedia, legislation, newspapers, books, web content, etc. The more recently published texts are still copyright-restricted, which limits the availability of the full corpus. The NCC has texts in both written languages, but the Nynorsk proportion is significantly smaller (5-10%). To remedy the scarcity of Nynorsk text data, the Language Bank at the National Library harvests available legal documents from municipalities where Nynorsk is the main language.

There are three large language models (NorELMo, NorBERT, and Notram) for Norwegian, which have been trained on (parts of) the NCC. These models can be fine-tuned with annotated corpora to develop task-specific tools. The language models' embeddings are significantly less robust for Nynorsk than for Bokmål, again due to the disproportionate distribution of the languages in the training data.

Norway does not have access to the same amount of parallel data from the European institutions as the EU Member States. Even so, the ELRC initiative, in which Norway participates, has contributed to a growing awareness of the reusability of translations. Public administrations have contributed significant collections of Bokmål-English parallel data. Valuable translation memories for developing MT systems from English to Bokmål have also come out of EU-funded research projects, e. g., PRINCIPLE. There are very few translation memories between Nynorsk and English, but it is possible to use Bokmål as a pivot language when developing MT technology for English-Nynorsk. The most prominent Nynorsk-Bokmål corpus is the manually corrected output of the Nynorsk press agency Nynorsk Pressekontor's Apertium-based pipeline. Due to the similarities between Nynorsk and Bokmål, MT between the two written standards yields fairly good results.

The most important lexical resource for Norwegian is Norsk ordbank (the Norwegian Word Bank), a lexical database for Norwegian Bokmål and Nynorsk reflecting the official standard orthography as defined in the Norwegian dictionaries Bokmålsordboka and Nynorskordboka. Both are freely available for download and use in LT. While some domain-specific termbases exist for Bokmål, very few terms appear in their Nynorsk parallel, for instance in the national terminology portal Termportalen.

While there is no research programme in Norway aimed specifically at LT, several projects are in the process of filling some of the identified gaps in Norwegian LT and LRs. All major universities in Norway conduct research on LT and/or AI. Among the running projects, NorwAI aims at developing LTs for Scandinavian languages, including conversational search in natural language. SCRIBE seeks to develop an advanced speech-to-text transcription system for spontaneous speech. SANT (Sentiment Analysis for Norwegian Text) is to create open LRs for sentiment analysis for Norwegian. The public broadcasting corporation NRK and two private media groups contribute to the project. The Målfrid project collects all available digital texts from the public sector in Norway. An effort like this will ensure the availability of unstructured text data of a more recent date. CLEANUP aims to develop tools and techniques to automatically anonymise unstructured text data from an array of domains. The project Universal Natural Language Understanding builds upon the UD standard for syntactic treebanks. The goal of the project is to convert the syntactic representation to machine-readable semantic representation.

3 Recommendations and Next Steps

Even though the increase in data availability from 2018 to 2021 has been substantial, awareness of what language data is, what it can be used for and how it should be shared, needs to be raised in all sectors. Due to the lack of Nynorsk data and modern LTs' preference for big data, it must be a priority for decision makers to strengthen LT for the lesser-used language to avoid weakening its equal status. Public sectors must take on their new responsibility as required in the new language act and ensure parallel versions of Bokmål and Nynorsk LT in public procurement.

While there are certain synergies when developing parallel LT for both languages, there is also a need for parallel development of basic resources. The creation of missing tools and LRs must continue. There is a need for more text data for Nynorsk, more domain-specific data, and lexical/terminological resources, in particular for Nynorsk, as well as speech data that cover dialects and Nynorsk in addition to tools for semantic parsing. As for the quality of Norwegian LT, no overreaching assessment has been made of the improvement we assume has taken place. In particular, downstream (user-driven) quality assessment of Norwegian Nynorsk and Bokmål LT is needed, to compare their quality, as well as dialect understanding.

Political action is necessary to open up international platforms to include the possibility of introducing LTs for smaller languages such as Norwegian Nynorsk, even when the large platforms themselves do not offer LT for these smaller languages.

There must be sufficient funding for research and development for Bokmål and Nynorsk LT, and the extra cost of developing parallel versions of Bokmål and Nynorsk LT should be considered when funding future research programmes. A dedicated programme for LT should be considered. Participation in international research projects and programmes that focus on LT, should be encouraged.

References

Eide, Kristine, Andre Kåsen, and Ingerid Løyning Dale (2022). *Deliverable D1.26 Report on the Norwegian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-norwegian.pdf.

Smedt, Koenraad De, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard (2012a). *Norsk i den digitale tidsalderen (bokmålsversjon) – The Norwegian Language in the Digital Age (Bokmål Version)*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/norwegian-bokmaal.

Smedt, Koenraad De, Gunn Inger Lyse, Anje Müller Gjesdal, and Gyri S. Losnegaard (2012b). *Norsk i den digitale tidsalderen (nynorskversjon) – The Norwegian Language in the Digital Age (Nynorsk Version)*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/norwegian-nynorsk.