Maltese at the digital crossroads
(or: Some remarks on doing NLP for a “small” and “low-resourced” language)

Albert Gatt
Institute of Linguistics & Language Technology
University of Malta
albert.gatt@um.edu.mt
http://staff.um.edu.mt/albert.gatt
Current NLP Landscape

A growing interest in multilingual applications and variation.

How broadly is multilinguality construed?

Who benefits?

Transfer learning, pretraining, fine-tuning

Unsupervised learning

Variation (Social, geographical, temporal)
Commercial applications

Currently limited to a handful of languages, but a huge potential growth area.

Baidu Research Announces Breakthrough in Simultaneous Translation

Today, we are excited to announce STACL (Simultaneous Translation with Anticipation and Controllable Latency), the first simultaneous machine translation system with anticipation capabilities and controllable latency. It is an automated system that is able to conduct high quality translation concurrently between two languages. STACL represents a major breakthrough in natural language processing due in large part to the challenges presented by word order differences between the source and target languages, and the latency requirements in real-world applications of simultaneous translation or interpretation.

Baidu started a business 18 years ago and caught up with the Internet. In 2000, there were only 10 million Internet users in China.
Where will this go?

It’s never a good idea to predict the future destination of a technology.

But, from where we currently stand, there might be two plausible outcomes:

1. A handful of widely-spoken languages win out
   → NLP reinforces a state of affairs

2. We broaden the scope of multilinguality to its fullest extent.
   → NLP helps to challenge the state of affairs.
And policy should be our ally

... multilingualism presents one of the greatest assets of cultural diversity in Europe and, at the same time, one of the most significant challenges for the creation of a truly integrated EU.
This rest of this talk ...

The underlying thinking
NLP in a multilingual setting, including
• Languages with small(er) speaker populations
• Languages which are under-resourced

The specific focus
Maltese, with reference to:
• Historical/social aspects
• Challenges in developing resources and tools
• Technical challenges arising from the language itself

Nancy, France, 2019
Outline

1. Overview of the linguistic situation in Malta
   – Challenges for Maltese NLP
   – Is Maltese under-resourced?

2. Case Study #1: Hybrid morphology and automatic labelling

3. Case Study #2: Developing ASR with low resources

4. Some conclusions
Outline

1. Overview of the linguistic situation in Malta
   – Challenges for Maltese NLP
   – Is Maltese under-resourced?

2. Case Study #1: Hybrid morphology and automatic labelling

3. Case Study #2: Developing ASR with low resources

4. Some conclusions
Some facts

Source: NSO, Census 2011
Current population probably > 500k
Linguistic history

Origins
Classical Arabic & Siculo-Arabic
870 - 1090

Contact
Arabic → Maltese
Romance influx
1090 - 1530

Mixture
Maltese
Prestige langs: IT, EN
1530 – 1798
1800 - 1964
Current linguistic situation

Malta has 3 official languages

• Maltese
• English
• Maltese Sign Language (as of 2016)

Maltese & English

• Standard Maltese with dialects
• Good evidence for a Maltese English variety (e.g. Grech, 2015; Bonnici, 2010)
  – Significant minority of native M. Eng speakers.
Current Linguistic Situation

Italian & Arabic

Census, 2011

Not at all
A little
Average
Well

Arabic

Italian

> 89  80-89  70-79  60-69  50-59  40-49  30-39  20-29  10-19

Nancy, France, 2019
Current Linguistic Situation

Spoken vs written modalities

• Speech:
  – Majority speak Maltese, with English as a second language. But varying degrees of bilingualism.

• Writing:
  – Vibrant literature in Maltese.
  – Strong preference for everyday written communication in English.
  – Stronger presence of EN on social media, more support for writing in EN.
  – Everyday writing in Maltese often “noisy” (esp. on social media)

Contact effects

• Heavy lexical borrowing (EN to MT)
• Code-switching (MT – EN)
IS MALTESE UNDER-RESOURCED?
Rosner and Joachimsen (2012) *METANET white paper*

- Comparison of 30 European Languages
- On all language technologies and resources, support for MT ranked at weak/no support

Remember those alternatives

1. A handful of widely-spoken languages win out
   → NLP reinforces a state of affairs

2. We broaden the scope of multilinguality to its fullest extent.
   → NLP helps to challenge the state of affairs.

It is often easier for users to avoid using MT in digital contexts. The support for EN is simply greater.
What is an under-resourced language?

Definitions are often, implicitly **relative**.

- BLARK (Krauwer, 2004)
- Other definitions (e.g. Berment, 2004; Besacier et al, 2014)

**Example:**
While counting languages is a tricky task, the number of “well-resourced languages” can be easily given by listing how many languages are identified for core technologies and resources, such as: Google Translate [...], Google search [...], Siri ASR application [...], Wiktionary [...], Google Voice Search [...] (Besacier et al, 2014)
What is an under-resourced language?

It depends...

A language is under-resourced with respect to NLP Task X if at least one of the following holds:

• Core aspects of the language at the relevant level are under-studied or not standardized.
  – E.g. ASR for “small” languages whose phonology is poorly understood
  – E.g. The language is unwritten

• In one or more relevant modalities, there is no standardisation
  – E.g. There is no standardized orthography

• Data to train models in Task X is lacking
  – E.g. Speech-text pairs for ASR
  – E.g. Digital lexical resources for morphological labelling
  – E.g. Parallel or comparable corpora for MT

• Core supporting technologies are lacking
  – E.g. POS Tagging to support (some forms of) parsing
Server għar-Riżorsi Lingwistichi Maltin

L-MLRS huwa proġett li jiffoka fuq il-holqien ta' riżorsi u ghodod għall-lingwi tal-Gżejer Maltin, jiġifieri l-Matti, l-Ingliż u l-Lingwa tas-Sinjali Mattija.

Riżorsi u ghodod

Ir-riżorsi disponibbli jinsabu f'dawn il-kategoriji:

- Korpora: korpora ġenerali ta' testi bil-Matti, kif ukoll korpora għall-edukazzjoni bl-Ingliż Malti u bil-Matti;
- Dizzjunari elettroniċi, li jinkludu dizzjunaru Ingliż-Matti u dizzjonaru tal-Matti Storiku;

Għajnuna

Il-proġett MLRS huwa kiordinat mill-Istitut tal-Lingwistika u d-Dipartiment tas-Sisemi Intelligenti tal-Kompjuter fi ħdan l-Universita ta' Malta.

Ibbenefika minn għajnuna finanzjarja mingħand:

- Il-Kunsill Malti għax-Xjenza u t-Teknoloġija (MCST);
- Il-Fond ta-Riċerka tal-Universita ta' Malta.

Ahbarijiet

Tnedija ta' Korpus ġdid

Inħabbru t-tnediija tal-Korpus Malti v3.0 (2016), verzjoni ġdida ta' korpus Malti. Dai-korpus fih madwar 250 miljun token, f'diversi kategorijiet, b'ittaggjar grammatikali ikktar akkurat, ġematizzazzjoni, kif ukoll l-annotazzjoni tal-gheruq morologiċi tal-klim.

Dizzjunaru Malti onlajn

The resource landscape

Lexical

• Ġabra
  – Full-form lexical DB
  – 17k entries; 4.5 million wordforms
  – EN glosses

• Ġabra tal-Malti Qadim
  – Historical lexicon (16-17 century)

• Dizzjunarju tal-Lingwa tas-Sinjali
  – LSM Dictionary
  – Multimodal
Example lexical entry

Core Procedure:
- Automatic (rule-based) generation
- Manual fine-tuning

Then:
- Crowdsourced user suggestions
- Harvesting from online resources
The resource landscape

**Lexical**
- Ġabra
  - Full-form lexical DB
  - 17k entries; 4.5 million wordforms
  - EN glosses
- Ġabra tal-Malti Qadim
  - Historical lexicon (16-17 century)
- Dizzjunarju tal-Lingwa tas-Sinjali
  - LSM Dictionary
  - Multimodal

**Corpora**
- Korpus Malti v3.0 (2016)
  - 250m tokens
  - Morphosyntactic annotation
  - Multi-genre
- Learner corpora (MT and EN)
  - Ca. 3m tokens
- MUDT – Universal Dependencies corpus (Ceplo, 2018)
  - 44k tokens
- (Other resources including Europarl, EU Translation memories.)
Tools

• POS Tagger (SVM):
  – Ca. 96-97% accuracy

• Morphological analysis/generation:
  – Stemming (Tanti, 2013)
  – Clustering and labelling of words (Borg, 2016, Ravishankar et al 2017)
  – Generation (Camilleri, 2013) and subsequent work

• Parsing
  – Align & transfer (Tiedemann and van der Plas, 2017)
  – Neural, low-resource approach (Zammit, 2018)
So what challenges for NLP does Maltese present?

**Linguistic**

- Long history of language contact gives rise to interesting grammatical and lexical complexities.
- Mixture of Arabic/Semitic and Romance elements permeates the grammar.

**Social**

- MT vs EN imbalance across written/spoken modalities in everyday communication
- Code-switching
- Bilingualism means the public relies on EN resources, making MT resource development appear less of a priority.
Outline

1. Overview of the linguistic situation in Malta
   – Challenges for Maltese NLP
   – Is Maltese under-resourced?

2. Case Study #1: Hybrid morphology and automatic labelling

3. Case Study #2: Developing ASR with low resources

4. Some conclusions

Nancy, France, 2019
LEARNING MORPHOLOGY

Work in collab. with Claudia Borg, Ray Fabri, Manolo Perea
A hybrid system

- Mixture of concatenative (Romance) and non-concatenative (Arabic) derivational systems.

<table>
<thead>
<tr>
<th></th>
<th>Derivation</th>
<th>Inflection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>gideb</strong> ‘to lie’</td>
<td>giddieb ‘liar’</td>
<td>giddieb-a (sg.f)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>eżamina</strong> ‘to examine’</td>
<td>eżaminatur ‘examiner’</td>
<td>eżaminatr-iči (sg.f)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NLP
ADVISORY
(Psycholinguistic Digression)
To what extent is Maltese morphology semitic?

• Romance derivation is highly productive (Hoberman & Aronoff, 2003; Mifsud, 1995; Gatt & Fabri, 2018).
  – Only about 1900 extant roots, most occur in 2 patterns on average.
  – Root-based morphology is semantically opaque.
The psycholinguistics

Rapid Serial Visual Presentation
The psycholinguistics

Rapid Serial Visual Presentation

*The journey was cancelled...*

jounrey cacnelled

See Velan & Frost (2007,...) and much subsequent work.
The psycholinguistics

Rapid Serial Visual Presentation

Marija libset id-dublett l-iswed

lisbet  iwsed

Root consonants are not an informative signal for recognition.
See Perea et al (2012)...

Nancy, France, 2019
The psycholinguistics

Rapid Serial Visual Presentation

Marija libset id-dublett l-iswed

Root consonants are not an informative signal for recognition. See Perea et al (2012)...

Root consonants are an important organisational element in the lexicon. Twist, 2006; Ussishkin & Twist, 2009; Ussishkin et al, 2015

Nancy, France, 2019
NLP ADVISORY
End of Digression
The upshot

• Dealing computationally with MT morphology requires us to deal with multiple co-existing systems.

• Different features matter: root vs stem, vowel pattern...
  – *kelma* VKLM ‘word’
  – *kalma* kalm-a ‘calm’
Unsupervised Clustering

- Borg & Gatt (2014 ...):
  - Identify affixes based on transitional probabilities
  - Cluster based on character and distributional semantic similarity.

- Evaluation using crowdsourcing and with experts (linguists; n=3).
  - How many clusters are modified?
Labelling morphological features

- **Strategy:**
  - Treat each feature/label as a classification problem
  - Identify the optimal classifier cascade for each POS
  - Core features: extracted directly from word form (prefix, suffix, cons, vow etc)
  - Cascade features: Classifier $C_{t+1}$ incorporates features identified from $(C_0 \ldots C_{t+1})$
  - Multiple classification algorithms (LR, Decision Tree, Random Forest...)
  - Data from Ġabra ('silver' - some automatically generated)
    - Train 170k, Test 20k
    - Held out ‘gold’ dataset of 200 words, manually annotated
Which features?

Verb example: (ma) ktibtulux

Tense/Aspect = perf
Gen = neut
Num = sg
Dir: 3SgM
Ind: 3SgM
Pol: neg

ktibt- u- lu- x
write.1Sg.perf 3SgM 3SgM Neg

“I didn’t write it for him"
Labelling morphological features (DT Example)

- “Traditional” = based on 10-fold cross-validation
- Gold Standard = against the held-out manually annotated set
Neural model

Main idea: avoid feature engineering, exploit sub-word regularities

\[ mort \]

- Char embedding
- BiLSTM Encoder
- Fully connected layer + Softmax

\[ Neg = 0, Asp = Perf, Person = 1, Number = Sg, Gen = Neut \]
Example results (verbs)

- Much improved performance on most features.
- Exploitation of subword regularities benefits classification.
So what are the challenges here?

Mainly linguistic

• Hybrid system is a challenge for standard, feature-based approaches.
• Feature engineering (when needed) a major challenge.
• We find marked improvements when we avoid feature engineering and exploit subword sequences.
  – Implicitly, this is what the neural approach is doing.
Outline

1. Overview of the linguistic situation in Malta
   - Challenges for Maltese NLP
   - Is Maltese under-resourced?

2. Case Study #1: Hybrid morphology and automatic labelling

3. Case Study #2: Developing ASR with low resources

4. Some conclusions

Nancy, France, 2019
AUTOMATIC SPEECH RECOGNITION

Work in collab. with Carlos Mena, Andrea De Marco, Lonneke van der Plas, Claudia Borg

Nancy, France, 2019
Speech Technology for Maltese

Current

• Diphone-based Text-to-Speech (Borg et al, 2013)
  – Under re-development
• Grapheme-to-Phoneme transcription (rule-based)

Automatic Speech Recognition

• No existing system
MASRI: Maltese Automatic Speech Recognition

Ongoing project at UM, start 2019

Aims:
1. Develop robust ASR models
2. Explore cross-lingual techniques
MASRI: Data

• Data collection is our main bottleneck!
• Virtually no text-speech corpora of significant size.
• Hard to source opportunistically
  – Imbalance between spoken/written modalities across EN/MT does not help.

Current scenario:
• Manual annotation of existing speech resources (esp. European Parliament debates)
• Elicitation: pre-selection of sentences from MLRS Corpus. Close and far-field recording.
• Data augmentation techniques: Noisification (via noise superposition or spectrogram manipulation)
# MASRI: Data

<table>
<thead>
<tr>
<th>Type</th>
<th>Quality</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elicited</strong></td>
<td>Clean</td>
<td>Ca. 12 hrs</td>
</tr>
<tr>
<td>Relatively balanced for gender, region/accent</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Spontaneous</strong></td>
<td>Euro Parliament Interventions</td>
<td>Ca. 0.5 hr</td>
</tr>
<tr>
<td></td>
<td>Conversational data, multi-dialect</td>
<td>Ca. 5 hrs</td>
</tr>
</tbody>
</table>
MASRI: Data

Crowdsourcing initiative

- Mozilla CommonVoice initiative
- Localisation and adaptation for MT
- To be launched end of 2019
MASRI: modelling approaches

Dimensions

Data:
Clean vs Augmented (+noise)

Monolingual baselines – LOW RESOURCE
End-to-End DNN systems (CommonVoice2, Jasper)

Cross-lingual approaches - 1
Pre-Train: large datasets from related languages (Arabic, EN, IT)
Fine-tune: MT

Unsupervised training
Wav2Vec-like solution for unsupervised pre-training

Nancy, France, 2019
So what are the challenges here?

Data

• Significant data bottleneck
• Low availability of recorded speech + transcription

“Sociological” challenges

• Crowdsourcing with a very small population
• Perception of English as a “natural” communication code in the digital sphere
Outline

1. Overview of the linguistic situation in Malta
   – Challenges for Maltese NLP
   – Is Maltese under-resourced?

2. Case Study #1: Hybrid morphology and automatic labelling

3. Case Study #2: Developing ASR with low resources

4. Some conclusions
CONCLUSIONS
Remember those alternatives

1. A handful of widely-spoken languages win out
   \( \rightarrow \) NLP reinforces a state of affairs

2. We broaden the scope of multilinguality to its fullest extent.
   \( \rightarrow \) NLP helps to challenge the state of affairs.

It is often easier for users to avoid using MT in digital contexts. The support for EN is simply greater.

But a concerted research effort has begun to reshape that landscape.

Nancy, France, 2019
Broadening the scope of multilinguality to its fullest extent?

This means we:

1. ... pay attention to the structure and function of a variety of communication codes;

2. ... seek to understand them in their social and historical contexts;

3. ... develop the right level of support to ensure equal access to communication technologies.
Broadening the scope of multilinguality to its fullest extent?

This means we:

1. … pay attention to the structure and function of a variety of communication codes;
2. … seek to understand them in their social and historical contexts;
3. ... develop the right level of support to ensure equal access to communication technologies.

In other words, use NLP to avoid reinforcing an existing scenario.
Strategies for under-resourced languages?

• Address core technologies and resources
  – Corpora
  – Basic tasks such as morphology, POS, ...

• Exploit learning algorithms that leverage data from neighbouring languages.
  – Transfer and pre-training
  – Fine-tuning

• Influence policy (!)
Thanks