



Prof Claudia Borg
Associate Professor
Department of Artificial Intelligence
Faculty of ICT, University of Malta
Affiliate with the Institute of Linguistics and Language Technologies

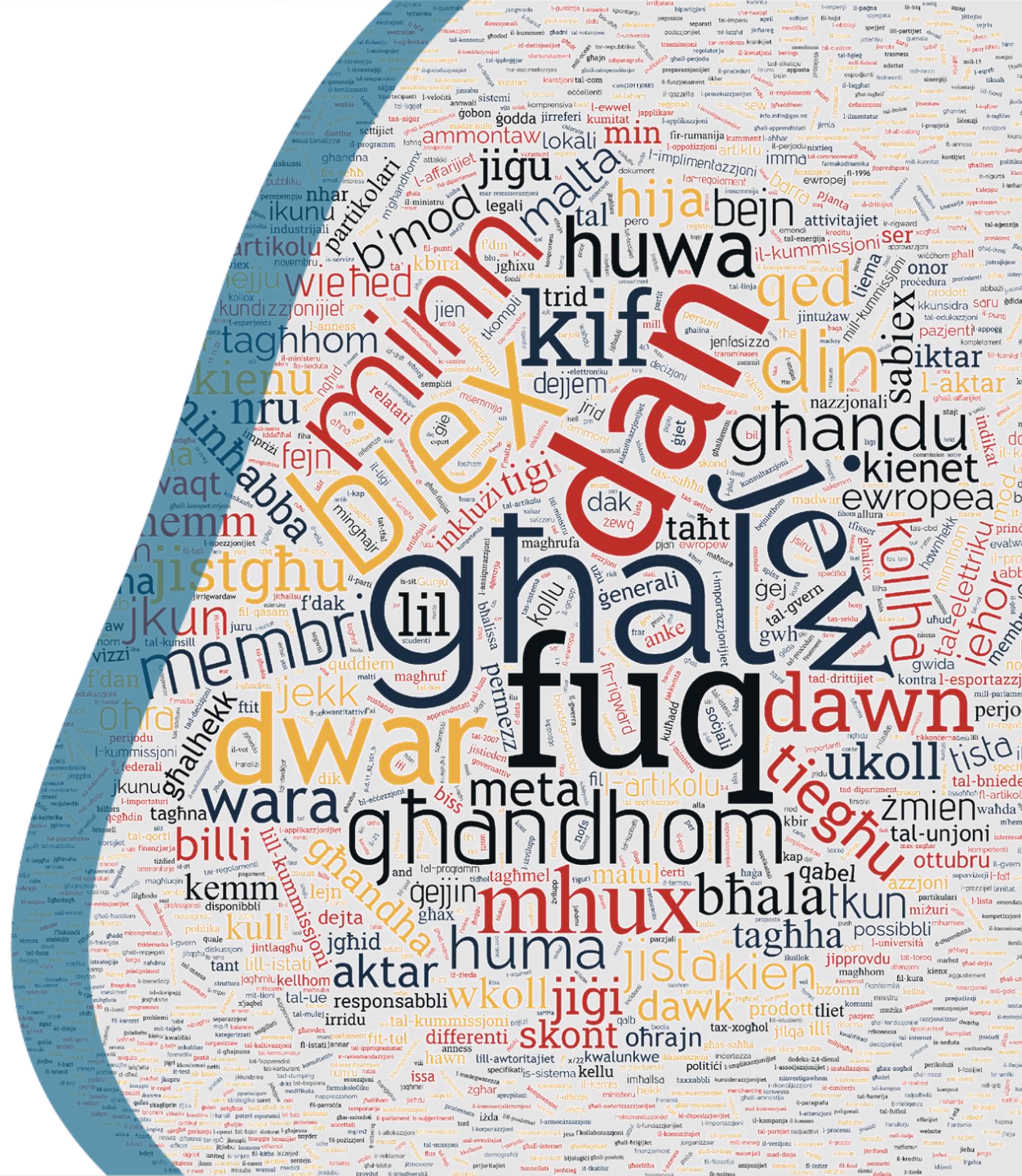
Associate Professor

Department of Artificial Intelligence

Faculty of ICT, University of Malta

Affiliate with the Institute of Linguistics and Language Technologies

L-Università ta' Malta



Korpus Malti v4.2 (KM)

- Contains ≈467M tokens
 - Continuation of Korpus Malti v3.0: roughly twice the size
 - Maltese Wikipedia is <1% of this corpus
- Collected from specific sources
 - Less noise
 - Richer metadata
- Split across 19 different domains/genres

subset	tokens	size
Wikipedia	1.89M	9.73MB
KM	467M	2.52GB

BERTu and mBERTu

Maltese Large Language Models

- Dependency Parsing
- Part-of-Speech Tagging
- Named-Entity Recognition
- Sentiment Analysis



BERTu



Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas and Claudia Borg.

[Pre-training Data Quality and Quantity for a low-resource language](#) in

Proceedings of the 3rd Workshop on Deep Learning for Low-Resource NLP. 2022

Maltese Dataset Summary

Previous Benchmark

Name	Task	train	validation	test
MLRS POS	Part-of-Speech Tagging	4,935	616	616
MUDT	Dependency Parsing	1,123	518	433
MAPA	Named-Entity Recognition	3,901	1,093	2,109
Sentiment	Sentiment Analysis	595	85	433

- Mostly structured tagging tasks
- Lack of semantic tasks; mostly low-level, syntactic
- Very small in size

Maltese Datasets Summary

Expanded Benchmark

Name	Task	train	validation	test
MLRS POS	Part-of-Speech Tagging	4,935	616	616
MUDT	Dependency Parsing	1,123	518	433
MAPA	Named-Entity Recognition	3,901	1,093	2,109
Sentiment	Sentiment Analysis	595	85	433
SIB-200	Topic Classification	701	99	204
Taxi1500	Topic Classification	860	106	111
News Tags ^[1]	(Multi-Label) Topic Classification	10,784	2,293	2,297
MultiEURLEX	(Multi-Label) Topic Classification	17,521	5,000	5,000
Belebele	Machine Reading Comprehension	0	0	900

Discriminative tasks

^[1][Topic Classification and Headline Generation for Maltese Using a Public News Corpus](#) (Chaudhary et al., LREC-COLING 2024)

Maltese Datasets Summary

Expanded Benchmark

Name	Task	train	validation	test	
MLRS POS	Part-of-Speech Tagging	4,935	616	616	Discriminative tasks
MUDT	Dependency Parsing	1,123	518	433	
MAPA	Named-Entity Recognition	3,901	1,093	2,109	
Sentiment	Sentiment Analysis	595	85	433	
SIB-200	Topic Classification	701	99	204	
Taxi1500	Topic Classification	860	106	111	
News Tags ^[1]	(Multi-Label) Topic Classification	10,784	2,293	2,297	
MultiEURLEX	(Multi-Label) Topic Classification	17,521	5,000	5,000	Generative tasks
Belebele	Machine Reading Comprehension	0	0	900	
MultiWOZ Maltese [†]	Task-Oriented Dialogue tasks	0	0	0	
OPUS100	Machine Translation	1,000,000	2,000	2,000	
FLoRes-200	Machine Translation	0	997	1,012	
WebNLG	Data to Text	*13,211	1665	1,778	
Eur-Lex-Sum	Abstractive Summarisation	940	187	188	
News Headlines ^[1]	Abstractive Summarisation	17,782	3,810	3,811	

*Indicates noisy data obtained through machine translation.

[†]Currently being constructed.

^[12][Topic Classification and Headline Generation for Maltese Using a Public News Corpus](#) (Chaudhary et al., LREC-COLING 2024)

Spell Checking

- An interface is being created to collect natural spelling errors for the Maltese language.
- The user is prompted to listen to a sentence in Maltese and type it out in a text box
- The backspace button is blocked out so that all typing errors are collected.
- Scan the QR code and give it a go!



Ġabra ta' Żbalji Ortografiċi

▶ 0:03 / 0:03 — 🔊 ⋮

Ittajpja dak li smajt hawnhekk:

Tista' tbiddel il-ħeffa tar-recording hawnhekk:

Normali ▼

Ibgħat

Frug

Machine Translation



- Worked on the EU Translation Platform
 - Can be accessed through **traduzzjoni.mt**
- Fixing public datasets that have inconsistencies (OPUS-100)



Machine Translation

- Future Research:
 - Expanding domain coverage of Machine Translation to Maltese
 - “Infusing” a machine translation system with a terminology database
 - Deal with complex morphology

National Language Technology Platform

Use the best language technology tools. [See all the tools](#)

English ▼	↔	Maltese (General) ▼
She drove it for <u>him</u>	×	Saqitha ghalih

Maltese ▼	↔	English (General) ▼
Saqithielu	×	She said it to him



OFFICE OF THE STATE ADVOCATE

Arabic-Maltese Transliteration

- Tools to transliterate Maltese text to Arabic script and vice versa
 - A dataset with etymology labels
 - A classifier to predict these labels
- Allow us to improve cross-lingual transfer from multilingual models
- Enables us to leverage data from related languages when building our language models

Arabic	Non-Arabic	Name	Arabic	Symbol	Arabic	Code-Switching	Non-Arabic
Il-	karozza	Porsche	tal-	2022	għandha	speed	fenomenali
ال	كردزة	برسكهي	تاع ال	٢٢٠٢	عندها	صباد	فنمنلي
ال	karozza	Porsche	تاع ال	2022	عندها	speed	fenomenali

Mailing list



Maltese Digital Resources Workshop 2025

Various new resources are being developed to integrate the Maltese language into the world of digital technology and artificial intelligence. Would you like to know more?

The second annual workshop will be held on the 11th of February 2025, 9:00 – 13:00, at the University of Malta (I-lmsida) and online on Zoom.

To attend, register at

<https://forms.gle/tkiuC6rC6e2qo6T7q>

Send your questions to Dr Marc Tanti at

marc.tanti@um.edu.mt

Programme:

08:30 – 09:00	Registration
09:00 – 09:10	Introduction
09:10 – 10:30	Presentations
10:30 – 11:00	Coffee break
11:00 – 12:20	Presentations
12:20 – 13:00	Open discussion



Current Directions

- Increasing datasets
- Initial experiments with further tuning of LLMs with a focus on Maltese
- Adding transliterated Arabic (into Maltese script) to increase cross-lingual transfer

Grazzi!

If you have any questions, now is the right time to ask!
Contact me on the details below for further info:

 Claudia Borg

 /AlMalta

 claudia.borg@um.edu.mt

