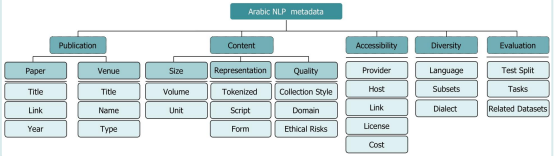
Masader: Metadata Sourcing for Arabic Text and Speech Data Resources

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb and Maged S. Al-shaibani

Masader is the largest public catalogue for Arabic NLP datasets, which consists of 200 datasets annotated with 25 attributes. , we develop a metadata annotation strategy that could be extended to other languages. We also make remarks and highlight some issues about the current status of Arabic NLP datasets and suggest recommendations to address them.

Metadata

25 attributes addressing different dimensions: publication, content, accessibility, diversity, and evaluation

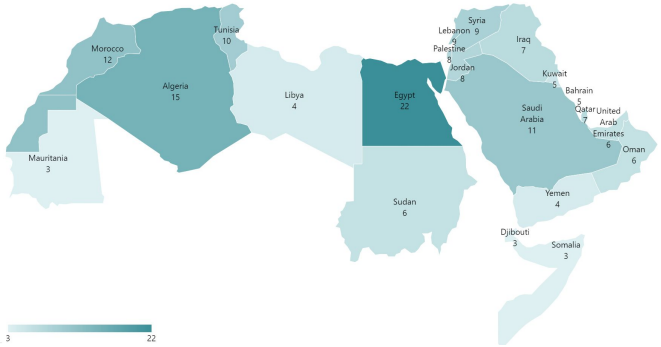


Statistics

Unit	Volume
Tokens	451,370,314
Sentences	1,236,350
Documents	51,701
Hours	3,104.1
# Datasets	200
# Datasets with Dialect Subsets	23
# Total Subsets	375

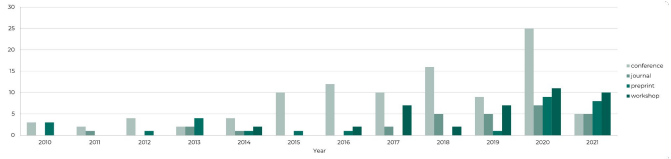
Dialect Resources

We also notice a shift in the dialect variety of published datasets in the Middle East and North Africa, as more datasets address low-resource dialects.

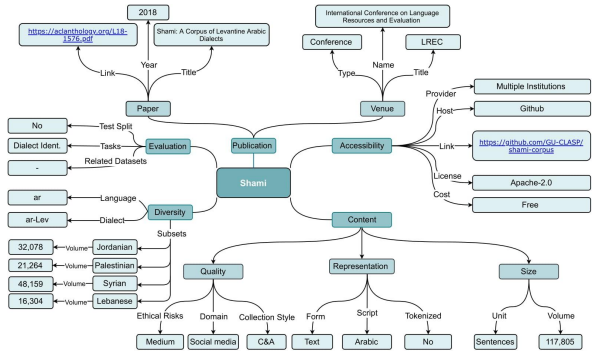


Arabic NLP Publications

We were able to offer a snapshot of the present state of Arabic NLP datasets in terms of publications using the annotated datasets. In this regard, we have seen a trend of publishing more datasets in different venues.



Sample Resource



Interface

Show 10 entries Search:

No.	Name	Link	Year	Volume	Unit	Paper Link	Access	Tasks
1	Shami	GitHub	2018	117,805	sentences	Shami: A Corpus of Levantine Arabic Dialects	Free	• dialect identification
2	LABR	GitHub	2013	63,257	sentences	LABR: A Large Scale Arabic Book Reviews Dataset	Free	• sentiment analysis
3	Arabic POS Dialect	GitHub	2018	1,400	sentences	Multi-Dialect Arabic POS Tagging: A CRF Approach	Free	• part of speech tagging

Final Remarks

- Masader was developed as part of the BigScience initiative 🌸
- Masader is on-going project. You can inspect the datasets and contribute using the Masader Interface