Available online at www.sciencedirect.com

ScienceDirect

00 (2017) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

A morphological analyser for Maltese

Vinit Ravishankar[a], Francis M. Tyers[b], Albert Gatt[c]

[a] *Faculty of Information and Communication Technology, University of Malta*
[b] *School of Linguistics, Higher School of Economics, Moscow, Russia*
[c] *Institute of Linguistics and Language Technology, University of Malta*

Abstract

This article describes the development of a free/open-source morphological description of Maltese, originally created as the analysis component in a rule-based machine translation system for Maltese to Arabic and later applied to other tasks. The lexicon formalism we use is *lttoolbox*, part of the Apertium machine translation platform. An evaluation of the analyser shows that the coverage is adequate, at 84.90%, while precision is 92.5% on a large automatically annotated test set and 96.2% on a smaller hand-validated set.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: finite-state, maltese, morphological analysis, apertium, templatic morphology, lttoolbox

1. Introduction

In this paper, we describe the development of a free/open-source morphological analyser for Maltese, a language closely related to north African dialects of Arabic, with around 520,000 native speakers, most of whom live on the island of Malta. A morphological analyser is a computational system that models a language's morphology, and is used to output morphological analyses from word forms and generate word forms from morphological analyses.

Maltese is a Semitic language spoken in the European nation of Malta. From a historical perspective, its origin is from Arabic, but Romance — especially Italian — and English have had substantial influence on its morphology [8, 13, 9, 17]. The Semitic component of Maltese utilises several archetypal Semitic grammatical patterns, like the triconsonantal root system; the morphology also has distinct morphological patterns for loanwords incorporated into the language. The extent to which the morphology can be characterised as Semitic is a matter of debate. For example, Żammit [19] sampled 1,820 Quranic Arabic roots and found that 40% of them were present in Maltese. Spagnol [17] listed all the extant roots in Maltese and found that most of them have significant paradigmatic gaps: of the 10 possible 'declensional' forms for verbs, for example, most roots exhibit only two. Furthermore, there has been significant descriptive work suggesting that the most productive component of the morphology is based on Romance affixation processes [13].

Section 2 gives a brief a description of Maltese morphology, and the features relevant to our analysis. Section 3 describes the *lttoolbox* paradigm system and how these paradigms are used to represent Maltese morphology. Section 5 is a description of our lexicon, along with a description of the tagsets used. Section 6 is an evaluation of both coverage on two corpora, and precision and recall on two data sets. Section 7 gives some perspectives for future development.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

2. Morphology

2.1. Nominal

Maltese nouns can be either masculine or feminine, and feature a dual — albeit with a very limited domain — along with the singular and the plural. Dual formation is concatenative, and adds the suffix *-ejn*, whilst also applying morphophonemic rules like vowel assimilation and elision at stem endings. Plurals are more irregular, and form either concatenatively — by adding one of several suffixes — or non-concatenatively, with irregular stem alterations (*broken plurals*). Words of non-Semitic origin typically pluralise concatenatively, with suffixes distinct from the Semitic-origin words that also pluralise concatenatively. Nominals are not marked for grammatical case.

Possessive enclitic pronouns can append to nouns to mark the gender, number and person of the possessed object. In the case of the dual possessor, the final *-n* is omitted. Such enclitic pronouns are also constrained in their domain of application, largely to a set of inalienable nouns (especially kinship and bodypart terms), together with some other nouns, for example: *omm-u* 'his mother'.

2.2. Articles

The definite Maltese article is *l*, which is orthographically connected to the following noun or adjective with a hyphen, and phonologically a proclitic. When appended to a word that begins with a consonant, the article begins with a vowel, as in *il-*. When followed by one of the *sun consonants*,[1] the article's consonant assimilates to the first consonant of the word it is attached to, e.g. *is-sistema* 'the system'. Note this orthographic choice makes analysing the Maltese article simpler and less ambiguous than the Arabic equivalent النِّظَام *alniẓām* 'the system'.

2.3. Verbal

Maltese verbs feature the triconsonantal root system, and have non-concatenative morphology, similar to most other Semitic languages. This is non-trivial to represent with the *lttoolbox* format, which is more optimal for suffixing, concatenative languages. Non-concatenative formation only marks tense/aspect/mood and subject gender, number and person: polarity and object inflection are marked by using either a circumfix for negation, or by appending pronominal suffixes, which are both easily represented with the *lttoolbox* paradigm system. The verb in Maltese can in fact take the same set of enclitic pronouns available for expressing possession in nouns, but here functioning to mark the direct object, for example, *serqit-u* 'she robbed him'. Indirect objects are also marked with these pronouns, but usually require the prefixation of *l*, probably a short form of *lil*, which is a differential object marker in Maltese. Thus: *serqit-u-li* 'she stole it from him'.

One difference between nouns and verbs where enclitic pronouns are concerned is that in the former, the pronoun marking first person singular is *-i*, as in *omm-i* 'my mother', whereas it is *-ni* that marks the object on verbs: *seraq-ni* 'he robbed me'.

The negation circumfix is not strictly a single affix: verbs take *-x* as a suffix, and the adverb *ma* is inserted before the verb, similar to the *ne - pas* construction in French.

There are five main conjugation classes for Maltese verbs of Semitic origin: strong verbs, that have triconsonantal or quadrilateral roots, defective verbs, that have a silent third radical (orthographically represented as *għ*), weak verbs, that have a semivowel for the third radical (*j/w*), hollow verbs, that have long vowels between the initial and final radicals (eg. *DaM* 'he was delayed'), and doubled verbs, with identical second and third radicals.

Maltese verbs, similar to other Semitic verbs, can undergo a variety of transformations via affixes, that alter the meaning of the root, eg. to introduce causativity. There are ten such transformations; however, no verb demonstrates all ten, as noted by Spagnol [17] and discussed above.

[1] That is, the letters: 'ċ', 'd', 'n', 'r', 's', 't', 'x', 'ż', and 'z', corresponding to phonemes which are coronal.

```
forms['pp.m.sg'] += [('im' + r[0] + 'ie' + r[1] + 'e' + r[2] , '-', 'LR')]
forms['pp.f.sg'] = [('m' + r[0] + 'ie' + r[1] + r[2] + 'a' , '-', '-')]
forms['pp.f.sg'] += [('im' + r[0] + 'ie' + r[1] + r[2] + 'a' , '-', 'LR')]
forms['pp.mf.pl'] = [('m' + r[0] + 'e' + r[1] + r[2] + 'in' , '-', '-')]
forms['pp.mf.pl'] += [('im' + r[0] + 'e' + r[1] + r[2] + 'in' , '-', 'LR')]
```

Fig. 1: Example code for the generation of pattern 3 past participles for strong verbs. The `LR` 'left-to-right' allows analysis but not generation of a particular form.

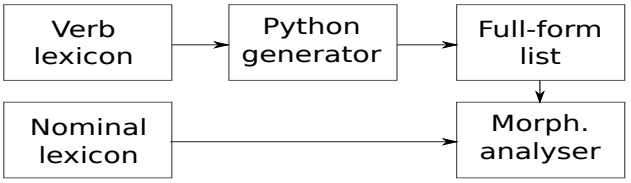

Fig. 2: Flowchart showing how the components of the analyser are put together

3. Paradigms

The *lttoolbox* paradigm system [15] describes finite-state transducers in XML. Paradigms are expressed as an input side (within '`<l></l>`' tags), and a corresponding output side (within '`<r></r>`' tags); the transducer is made to return the lemma of a word and the corresponding tags

Due to the fusional complexity of Maltese, we have used joins between paradigms to display analyses of separate morphemes separately. This essentially redirects the FST to another paradigm after it consumes the input for the first. The use of clitic pronouns is represented by *optionally* redirecting the FST to another paradigm. The paradigm system was not applied to internal verbal morphology; each verb form was, instead, a separate entry.

Other morphological analysis toolkits, such as HFST or Foma, would have allowed us to use replace rules to handle morphophonological alteration and templatic morphology, similar to Beesley's approach to Arabic morphological analysis [3]. Our justification for using *lttoolbox* despite the apparent inconvenience is twofold — firstly, it allows for rapid integration within the Apertium machine translation system [10], where it was designed for use in a Maltese-Arabic translation pair. Secondly, the use of Python scripts to generate templatic morphology, whilst seemingly fairly inelegant, was simpler to implement for the linguist who was already familiar with Python than, for instance, implementing Beesley and Karttunen's compile-replace rules [4] or the method described by Attia et al. [2].

In addition, by not using *flag diacritics*, we can export the transducer in table format such that it can be directly used by a wide range of other FST libraries, such as OpenFST [1].

3.1. Verbal inflection

It is non-trivial to represent templatic verbal morphology using the *lttoolbox* paradigm system for internal vowel changes; therefore, every form of a verb (excluding forms generated by suffixes) has been generated and entered as a separate entry into the analyser. The generation of these verb forms, particularly for the ten derived verb classes, was done by means of a Python script that would split up a verb stem into its constituent radicals, and incrementally build a full-form list by adding affixes wherever necessary. These would involve generated derived verb classes, as well as inflection for tense/aspect/mood, and gender/number/person (of the subject). Further verbal morphology, like negation and pronominal suffixes, were handled using *lttoolbox* paradigms.

Similar scripts were also used for the generation of paradigms for words of non-Semitic origin; these were relatively simpler as they do not follow the root system. Figure 2 is a block diagram of this system.

4. Prior work

To date, work on Maltese morphological analysis has been limited. The most comprehensive rule-based system was developed by Camilleri [7] as part of a computational grammar for Maltese based on the Grammatical Framework

Stem	Categ	Type	Pattern	Gloss	Root	Perf	Impf	Valency	pprs	pp
ġara	vblex	weak	1	run	ġ-r-j	a-a	i-i	iv		mi
refa	vblex	strong	1	raise	r-f-għ	e-a	i-a	tv		me
żied	vblex	hollow	1	increase	ż-j-d	ie-a		tv		mi

Stem	Categ	Type	Pattern	Gloss	Root	Infix	Impf	Valency	pp
kanta	vblex	loan	first_cons	sing	kant		a	tv	kantat
aċċetta	vblex	loan	first_vowel	accept	aċċett		a	tv	aċċett
ammira	vblex	loan	first_vowel	admire	ammir		a	tv	ammirat

Table 1: A sample of lexicon entries from the verb lexica. There are a total of 469 entries. The top table shows entries from the lexicon of Semitic verbs, while the bottom table shows entries from the lexicon of loan verbs.

(GF) [16]. The output of this analyser-generator has since been incorporated into a large online lexical resource, called Ġabra.[2] However, Camilleri's work was primarily focused on verb inflection and derivation, with the system generating the full inflectional paradigm for a verb based on the roots and patterns specified in the work of Spagnol [17]. The system also handles pronominal suffixes for verbs, though these are not part of the GF linearisation table, but as separately handled.

Borg [5] presented an in-depth investigation of machine learning approaches to Maltese morphology, using both unsupervised techniques for clustering morphologically related words [6] and supervised classifier cascades for labelling inflectional and pronominal features. One of the challenges noted by Borg is that the hybrid nature of Maltese morphology may compromise the performance of a 'one size fits all' solution, since some techniques can work better on lexical items generated on the basis of stem and affix morphology (from Romance, in the case of Maltese), while template-based morphology involving a root and pattern may benefit from different techniques, since here, crucial morphological information is incorporated in a discontinuous sequence of consonantal radicals and vowel melodies.

To date, there has been no analysis of the coverage or precision of a Maltese morphological analyser against a corpus of naturally occurring texts.

5. Lexicon

5.1. Tagsets

The native tagset of the analyser is based on the conventions of the Apertium project [10]. This follows from its development as part of the development of machine-translation systems for Maltese–Hebrew and Maltese–Arabic.

In addition, we provide a mapping to the part-of-speech and morphological standards of the Universal Dependencies project [14]. Figure 3 shows an example sentence in the Universal Dependencies format, with the relevant columns left in: the last two columns indicate the universal part-of-speech tag and UD-style morphological features, derived by converting our Apertium analyses (see §A.6).

5.2. Creation

Creating our lexicon involved manually adding entries from a frequency list. We generated this frequency list from a dump of the Maltese Wikipedia. Whilst adding lexical entries, when we came across unanalysed tokens, we added the entire paradigm for the token, and not just the surface form. This led to rapid increases in coverage. Table 2 is a brief summary of the number of paradigms per part of speech, and the number of forms that they generate. In addition to the XML lexicon entries, we also have a text-based system for the verbs, example entries can be found in Table 1.

[2] `http://mlrs.research.um.edu.mt/resources/gabra/`

Category	Paradigms	Entries	Forms
Verb*	26	469	484,638
Proper noun	11	3,770	3,770
Noun	544	2,998	46,233
Adjective	81	994	2,490
Adverb	8	183	266
Numeral	17	89	136
Determiner	11	35	287
Preposition	5	123	819
Pronoun	19	63	157
Conjunction	3	45	27
Interjection	1	14	14
Total:	726	8,779	538,837

Table 2: The total number of lexemes categorised by part of speech. * The number of paradigms for verbs is based on the number of stem types (e.g. hollow, doubled, quad, …) and the number of verb classes (e.g. 1, 2, 3a, 3b, …). Enclitic pronouns are not included in this count.

```
1      Matul       matul       ADP     _
2      l-          l           DET     Definite=Def|PronType=Art
3      istorja     storja      NOUN    Gender=Fem|Number=Sing
4      ,           ,           PUNCT   _
5      il-         l           DET     Definite=Def|PronType=Art
6      pożizzjoni  pożizzjoni  NOUN    Gender=Fem|Number=Sing
7      ta'         ta'         ADP     _
8      Malta       Malta       PROPN   Gender=Fem|Number=Sing
9-10   fil-        _           _       _
9      fi          fi          ADP     _
10     l-          l           DET     Definite=Def|PronType=Art
11     Baħar       baħar       NOUN    Gender=Masc|Number=Sing
12     Mediterran  Mediterran   PROPN   Gender=Fem|Number=Sing
13     kellha      kellu       AUX     Gender=Fem|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
14     sinjifikat  sinjifikat   NOUN    Gender=Masc|Number=Sing
15     strateġiku  strateġiku   ADJ     Gender=Masc|Number=Sing
16     .           .           PUNCT   _
```

Fig. 3: Example output of the analyser for the UD tagset. The translation of the sentence is 'Throughout history, the position of Malta in the Mediterranean Sea has had strategic significance'. The line 9–10 shows two level tokens where there are two underlying syntactic words for one surface form.

6. Evaluation

6.1. Quantitative

We evaluated our morphological analyser on two corpora: the entire Maltese Wikpedia, and the Korpus Malti. [11][3] Initially, we evaluated naïve coverage by calculating the percentage of surface forms that received at least one morphological analysis. Table 3 describes the corpora we used for the naive tests.

[3] `http://mlrs.research.um.edu.mt/`

Corpus	Tokens	Coverage (%)
Wikipedia	1.64M	85.00
MLRS	241.3M	84.80
Average	–	84.90

Table 3: Corpora used for naïve coverage tests

Corpus	Known tokens		All tokens	
	P	*R*	*P*	*R*
Automatic	92.5	–	77.3	–
Hand validated	96.2	95.3	–	–

Table 4: Percentage of tokens in the Korpus Malti which had tags which were found in the output of the analyser.

An issue with the Maltese Wikipedia was the existence of sections in Italian, often complete sentences. We used `langid.py` [12] to pre-process the Wikipedia corpus, and filter out sentences that were parsed as Italian; this reduced the token count in the corpus from 1.704M to 1.643M.

The Korpus Malti is annotated with POS tags; whilst this would not have been sufficient to evaluate complete morphological analysis, we mapped the POS tags used in the corpus to Apertium's standardised POS tagset. The two tagsets have a many-to-many relationship. Whilst some of the reductions in descriptiveness involve reducing subclasses of nouns and pronouns to a single category, others involve mapping particles (including the focus, future and negation particles) to adverbs in the Apertium tagset.

We then calculated how many tokens had at least one Apertium POS tag, thereby ignoring morphological ambiguity, in common with at least one converted Korpus Malti POS tag. As we did not run our morphological analyser on a stream of running text, instead running it on each word type in the corpus, the coverage figures differ over here: our tokenisation standards are different to the ones used in the corpus.

Further, we also carried out a more fine-grained manual evaluation of the full morphological analysis of 250 unique tokens, that received at least one analysis with our analyser. Incorrect analyses were removed, and missing analyses were added to the Apertium output.

Our final results are presented in Table 4; the *all tokens* field is blank for our hand validated set as we only considered tokens that received at least one analysis. Our justification for this is that we had already calculated raw coverage; the accuracy of the analyses themselves was important to us here.

6.2. Qualitative

Sorting the tokens missing from our analyser by POS tags helps determine precisely what kinds of tokens are typically missing. The largest word class with missing tokens is, by far, common nouns; followed closely by proper nouns. The number of unanalysed nouns of either kind is more than 2.5 times the number of unanalysed verbs, which are the next frequently-missed word class.

Our manual evaluation was fairly robust; we provide a summary of the missing and incorrect analyses in Table 5. Amongst the errors (which showed some overlap), amongst Semitic verbs, 3 forms (1 lemma) had the incorrect lemma, 3 forms (2 lemmas) had lemmas absent in the dictionary[4], and 1 form was generated incorrectly by the Python script. 1 Semitic verb (*qabad*) also overgenerated 3 incorrect forms. Further, 2 nouns were parsed as adjectives, and 1 noun had the incorrect lemma. 1 Romance verb also failed to parse as a past participle, but did so as just an adjective.

[4] The reason these forms received any analyses at all is because our system "overgenerates" by assuming diacritics if none are given; for instance, *ingorru* was also treated as *inġorru*, which received an analysis.

Word type	Missing analyses	Incorrect analyses
Semitic verbs	7	8
Loan verbs	2	0
Nouns	4	1
Adjectives	0	1

Table 5: Missing and incorrect forms in morphological analyses

7. Future work

We have two immediate avenues for future work. The first is to expand the size of the lexicon. Although there exist other lexical resources, these may not contain all the information required for inclusion into the analyser. The Ġabra resource however provides a good candidate for incorporation into Apertium, as it contains rich morphological information in a full-form lexicon.

Another important avenue is to integrate our analyser into Maltese treebanks under the Universal Dependencies project [14]; our analyser could be used to enrich any potential work on a Maltese treebank within the UD framework with POS tags and fine-grained morphological features. Attempts have been made to bootstrap dependency parsers for Maltese [18]; the presence of fine-grained morphological information would help improve parsing results in similar future efforts.

8. Concluding remarks

In this paper we have presented the first wide-scale evaluation of a morphological model for Maltese. The model is a finite-state machine which is generated from a combination of affixation rules described in XML for all categories except verbs, and form-generation rules in Python for the verbs. The system shows reasonable coverage, in the mid-80% over two corpora. The precision and recall of the system measured on a manually evaluated test set are also satisfactory, at 96.2% and 95.3% respectively.

Acknowledgments

We would like to thank Maria Fronczak and Sagie Maoz for the bulk of development work on this transducer, and for the Google Summer of Code who supported it in the 2011 and 2012 editions. We thank Kevin Unhammer for helpful advice and the anonymous reviewers for their comments.

Appendix A. Tagsets

References

[1] Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., Mohri, M., 2007. Openfst: A general and efficient weighted finite-state transducer library, in: International Conference on Implementation and Application of Automata, Springer. pp. 11–23.

[2] Attia, M., Pecina, P., Toral, A., Van Genabith, J., 2014. A corpus-based finite-state morphological toolkit for contemporary Arabic. Journal of Logic and Computation 24, 455–472.

[3] Beesley, K.R., 1996. Arabic finite-state morphological analysis and generation, in: Proceedings of the 16th conference on Computational linguistics-Volume 1, Association for Computational Linguistics. pp. 89–94.

[4] Beesley, K.R., Karttunen, L., 2003. Finite-state morphology: Xerox tools and techniques. CSLI, Stanford .

[5] Borg, C., 2016. Morphology in the Maltese Language: A Computaional Approach. Ph.D. thesis. Institute of Linguistics, University of Malta. Malta.

[6] Borg, C., Gatt, A., 2014. Crowd-sourcing evaluation of automatically acquired, morphologically related word groupings, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).

[7] Camilleri, J.J., 2013. A computational grammar for Maltese, in: Proceedings of the 4th International Conference on Maltese Linguistics.

Tag	Description	Tag	Description
n	Noun	pl	Plural
vblex	Verb	sp	Singular / plural
vaux	Auxilliary verb	du	Dual
vmod	Modal verb	col	Collective
prn	Pronoun	p1	First person
det	Determiner	p2	Second person
np	Proper noun	p3	Third person
ij	Interjection	px1sg	First person singular possessive
pr	Preposition	px2sg	Second person singular possessive
rel	Relative	px3sg_m	Third person singular masculine possessive
adv	Adverb	px3sg_f	Third person singular feminine possessive
cnjadv	Adverbial conjunction	px1pl	First person plural possessive
abbr	Abbreviation	px2pl	Second person plural possessive
adj	Adjective	px3pl	Third person plural possessive
cnjcoo	Co-ordinating conjunction	tv	Transitive
cnjsub	Sub-ordinating conjunction	iv	Intransitive
recip	Reciprocal	pres	Present tense
dem	Demonstrative	past	Past tense
ord	Ordinal	inf	Infinitive
neg	Negative	pprs	Present participle
acr	Acronym	imp	Imperative
ref	Reflexive	pp	Past participle
an	Animate / inanimate	ger	Gerund
m	Masculine	subj	Subject
f	Feminine	obj	Object
nt	Neuter	def	Definite
mf	Masculine / feminine	ind	Indefinite
top	Toponym	qnt	Quantifier
org	Organisation	pos	Possessive
al	Other	itg	Interrogative
ant	Anthroponym	num	Numeral
cog	Cognomen	cm	Comma
comp	Comparative	sent	Sentence marker

Table A.6: Native tagset

[8] Drewes, A., 1994. Borrowing in Maltese, in: Bakker, P., Mous, M. (Eds.), Mixed Languages. 15 Case Studies in Language Intertwining. Ifott, Amsterdam, pp. 83–111.

[9] Fabri, R., 2010. Maltese. Revue Belge de Philologie et d'Histoire 88, 791–816.

[10] Forcada, M.L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Sánchez-Martínez, F., Ramírez-Sánchez, G., Tyers, F.M., 2011. Apertium: a free/open-source platform for rule-based machine translation. Machine Translation 25, 127–144.

[11] Gatt, A., Čéplö, S., 2013. Digital corpora and other electronic resources for Maltese. Corpus Linguistics 2013 , 96.

[12] Lui, M., Baldwin, T., 2012. langid. py: An off-the-shelf language identification tool, in: Proceedings of the ACL 2012 system demonstrations, Association for Computational Linguistics. pp. 25–30.

[13] Mifsud, M., 1995. The productivity of Arabic in Maltese, in: Cremona, J., Holes, C., Khan, G. (Eds.), Proceedings of the 2nd International Conference of l'Association Internationale pour la Dialectologie Arabe (AIDA), University Publications Centre, Cambridge, UK. pp. 151–160.

[14] Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D., 2016. Universal dependencies v1: A multilingual treebank collection, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

[15] Ortiz-Rojas, S., Forcada, M.L., Ramírez-Sánchez, G., 2005. Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. Procesamiento del lenguaje natural 35, 51–57.

[16] Ranta, A., 2011. Grammatical framework: programming with multilingual grammars. CSLI studies in computational linguistics, CSLI Publications, Center for the Study of Language and Information, Stanford (Calif.).

[17] Spagnol, M., 2013. A Tale of Two Morphologies: Verb Structure and Argument Alternations in Maltese. Phd thesis. University of Konstanz.

[18] Tiedemann, J., van der Plas, L., 2016. Bootstrapping a dependency parser for maltese .

[19] Żammit, M.R., 1998. Cognate roots in Qur'anic Arabic and Maltese, in: Proceedings of the Third International Conference of A.I.D.A.