ISO/TC 37/SC 4 **N 091**
ISO/TC 37/SC 4/WG 1 Nxx

ISO/TC 37/SC 4
Language resource management

2003-10-25

Annotating textual and speech data in Maltese

Compiled in response to the International Standards Organisation call for contributions towards the construction of a Morphosyntactic Annotation Framework

ISO TC 37/SC4 N047

Compiled by: Albert Gatt
With: Alexandra Vella
Joe Caruana

1. Introduction

The present document has been compiled in response to the call for contributions issued by the International Standards Organisation (ISO TC37/SC4 N047) towards the adoption of a morphosyntactic annotation framework. The document aims to contribute samples at the following levels, where the object language is Maltese:

a. *Tagging*: Specifically, part of speech tagging. A tagset for Maltese is included in §3. In addition, a number of problems that arise in relation to the morphosyntactic annotation of Maltese textual documents are described and exemplified, as are current solutions where available, in §2.
b. Annotation of transcribed speech. A small set of transcribed utterances are provided, on the basis of which some issues in their annotation are pointed out.

Our aim in the compilation of this document has been primarily to draw attention to linguistic phenomena that should be accounted for in a broad-coverage annotation scheme which aims to include the greatest possible number of languages.

2. Morphosyntactic (POS) tagging

To date, two alternative tagsets have been constructed for the annotation of Maltese textual documents. Unless otherwise stated, examples are given from one of them, described in Gatt (2001). A table with a full listing of categories and attributes in the tagset is given in §3. Both of the tagsets were constructed with reference to the guidelines laid out in the XML Corpus Encoding Standard (XCES) and more specifically, the XCES-compatible guidelines set out by the Expert Advisory Group on Language Engineering Standards (EAGLES). The main concern of this section is to illustrate some of the problems that arise from the annotation of Maltese texts, relative to specific aspects of its morphosyntax which the XCES/EAGLES annotation scheme does not address directly, and which frequently motivated extensions to the original scheme, or counter-intuitive modes of classification. Specifically, the following samples will illustrate:

a. tense and aspect marking in the verbal category;
b. the distinction between main, auxiliary and participial verbal forms;
c. the morphological amalgam used to mark negation in the verbal component;
d. the status of verb forms;
e. distinctions of number and degree in the nominal and adjectival categories;
f. the problem of distinguishing dependent particles, particularly enclitic pronouns and procliticised prepositions;
g. the status of case markers.

In what follows, the samples are divided according to category for ease of exposition.

2.1 Verbs

Tense and aspect
Main verbs in Maltese do not encode tense distinctions but are marked for aspect (perfective/imperfective) as shown in (1).

(1) a. Jimxi
Walk-3SgM-Ipf
'he walks' / 'he is walking'

b. mexa
walk-3SgM-Pf
'he walked'

The only verb that explicitly marks tense in Maltese is *kien* (be-3SgM[1]) 'to be'. While *kien* can function as a main verb, it can also precede the main verb and function as a tense (past/non-past) marker, as shown in (2). In the case of non-past, *kien* generally marks future (2b).

(2) a. kien mar
be-3SgM-PS go-3SgM-Pf
'he went' / 'he had gone'

b. ikun miexi
be-3SgM-NPS walk-3SgM-Ipf
'he will be walking'

In addition, *kien* has to be distinguished from aspect markers, a class of particles that precede the main verb in the imperfective and mark progressive (3a) and prospective (3b) aspect, both of which are distinguished from the default habitual interpretation of the imperfective.

(3) a. kien $\left\{\begin{matrix} qed \\ qieghed \end{matrix}\right\}$ jiekol
be-3SgM-PS PROG eat-3SgM-Ipf
'he was eating'

[1] Verbs in Maltese do not have an infinitival form; the citation form is the third person singular masculine.

b. kien $\left\{\begin{matrix} se \\ ha \\ ser \\ sejjer \end{matrix}\right\}$ jiekol

be-3SgM-PS PROS eat-3SgM-Ipf
'he was going to eat'

Aspect markers do not inflect for person, number or gender, with the exception of *qiegħed* and *sejjer*. Moreover, they have a different distribution from either main verbs or *kien*. Hence, a desideratum for a morphosyntactic annotation scheme is that it encode the distinction between main verbs, auxiliaries (*kien*) and aspect markers (*se, ħa*, etc), where:

a. the past/non-past distinction is carried by *kien*;
b. the present tense is the default case in the absence of *kien*, while the non-past encoded by *kien* is distinguished from present;
c. perfective/imperfective aspect is carried by the main verb;
d. the aspectual distinctions progressive/prospective are encoded by the aspect markers.

The desirability of making such distinctions has led to extensions of the original EAGLES framework. For instance, the tagset tags *kien* as auxiliary, but does not mark the distinction between present and non-past. Since aspectual distinctions in the EAGLES framework are subsumed by an *aspect* attribute, perfective/imperfective are its two values, leaving no room for further distinction between progressive and prospective under imperfective. Aspect markers had to be included under the *unique/unassigned* category. A verb phrase such as (4) is tagged as in (5).

(4) kien se jibda jiekol
be-3SgM-PS PROS start-3SgM-Ipf eat-3SgM-Ipf
'he was going to begin eating'

(5)

Phrase	Tags
kien	VA3SgMPs (auxiliary verb 'to be', 3rd pers, sing, masc, past)
se	UPV (unique/unassigned, pseudo-verb/aspect marker)
jibda	VV1SgMIpf (main verb, 1st pers, sing, masc, imperfective)
jiekol	VV1SgMIpf (main verb, 1st pers, sing, masc, imperfective)

Note that, in the tagged sample in (5), the prospective is unmarked, and the aspect marker *se* is tagged under the unique/unassigned category.

Participial forms

The active/passive distinction is carried by active or passive particles in Maltese. Participial verb forms are distinguished from main verbs in that they lack person inflection, although they do inflect for number, and gender in the singular. This is illustrated for active participles in (6).

(6) a. {jiena / inti / hu} nieżel
{I/you/ he} descend-SgM-ACT

b. {jiena / inti / hi} nieżla
{I/you/she} descend-SgF-ACT

c. {ahna / intom / huma} nieżlin
{we/you-Pl/they} descend-Pl-ACT

The tagset captured the active/passive distinction under the verbal attribute *voice*, as shown in (6) and (7).

(6) kienet imniżżla
be-3SgF-PS down-SgF-PASS
'{she/it-F} was down'

(7)

Example	Tag
Kienet	VA3SgFPs (auxiliary verb, 3rd pers, sing, fem, past)
imniżżla	VVPSgF (passive participle, sing, fem)

Note that the tagset does not adequately distinguish between main verbs (VV) and participles (VVP), making the participle distinct only through the voice attribute. The EAGLES framework has no 'participle' distinction under the verbal attribute *status*. The solution exemplified here is not ideal, since participial verb forms should be distinguished from main and auxiliary verbs, since they have a different distribution and morphosyntactic (inflectional) properties.

Negation

Negative marking on verbs is via a circumfix *ma –x*, as shown in (8a). On the other hand, the *–x* suffix can occur independently of the particle *ma*, particularly in questions of negative polarity (8b).

(8) a. *ma* mar -*x*
NEG go-3SgM-Pf -NEG
'he didn't go'

b. Mar -x il- baħar?
Go-3SgM-Pf -NEG DEF- sea
'He didn't go to the beach, did he?'

Hence, both *ma'* and *–x* need to be explicitly tagged as verbal negative markers, as shown in (9). In the tagset, negative markers were included (counter-intuitively) under the *unique/unassigned* category, since no separate category for negation was provided by the EAGLES scheme.

(9)

Word	Tag
ma	UMX 'verbal negator'
Mar*x*	VV3SgMPf+UMX 'main verb, 3rd sing, masc, perfective' + 'verbal negator'

Note the use of the + sign to indicate the morphological dependency of *–x* on its host verb.

Verb forms

Maltese verbs of Semitic origin have preserved elements of the root-and-pattern morphology found in language such as Arabic and Hebrew. While verbs with quadri-radical roots have two forms (or *binyanim*), triradicals have ten. A four-way contrast for the verb *kiser* 'to break' is exemplified in (10), where forms are indicated in Roman numerals.

(10)

a. kiser
break:I-3MSg
'he broke x' (transitive)

b. nkiser
break:VII-3MSg
'he/it broke' (passive)

c. kisser
break-II-3MSg
'he made *x* break' (causative)

d. tkisser
break:V-3MSg
'he/it broke him/itself'
(passive/reflexive/intensified)

The forms were not included in the tagset under consideration, although there is scope for the extension of the set for their subsequent inclusion. Although the forms are of limited productivity in the derivational morphology of present-day Maltese, their inclusion in an annotation system may be desirable for the following reasons.

a. For NLP tasks such as parsing, the forms provide an indication at first pass of the meaning of the verb and its syntactic behaviour. Thus, the verb *nkiser* in (10a) is passive, and this is immediately apparent from its morphological form;
b. Annotation at this level can also be utilized in extracting frequency and other statistical information of different morphological forms for inclusion in computational lexica and the study of productivity;
c. For purposes of information extraction and content modeling, the association of a particular root with a concept or meaning is facilitated by the inclusion of information related to different forms in a corpus.
d. Allowing for the inclusion of verb forms in an annotation scheme accommodates data from languages, such as Arabic, where these morphological permutations are far more productive than they are in Maltese.

2.2 Adnominal categories

The adnominal categories considered here are nouns, adjectives, pronouns and determiners.

Number
Number inflection on nouns in Maltese distinguishes singular, plural and dual. Additionally, a *singulative* category can be identified. Semantically, the latter has the function of individuating elements of a plural set with cardinality less than or equal to ten. Although the function is mainly semantic, use of the singulative can have reflexes in the morphosyntax. In addition, a collective can be identified whose 'massifying' function is distinct from that of the plural. The contrast is brought out in (11).

(11) a. basla waħda
onion-Sg one
'one onion' (singular)

b. żewġ basal
two onions-Pl
'two onions' (plural)

c. żewġ basliet
two onions-Sgv
'two onions' (singulative)

d. biċċtejn
piece-DU
'two pieces' (dual)

e. ħafna nies
many people-Col
'many people' (collective)

Hence, a *number* attribute for the category *noun* makes a five-way distinction, tagged as in (12).

(12)

Word	Tag
basla	NCSgF 'common noun, singular, feminine'
basal	NCPl 'common noun, plural'
basliet	NCSv 'common noun, singulative'
nies	NCCl 'common noun, collective'
biċċtejn	NCDu 'common noun, dual'

While including such fine-grained distinctions should be allowed for in the number system, it should not be assumed that we are recommending that they be made mandatory.

Dimension

In the tagset, an attribute *dimension* was introduced, making a two-way distinction between the *default* (unmarked) value and the *diminutive*, which receives overt morphological marking, as shown by the contrast in (13). The diminutive applies to both nouns and adjectives and could not be included under *number*, since diminutives have number distinctions. Including *diminutive* under *number* would therefore detract from the discriminative value of this attribute.

(13)

Word	Tag
dar	NCFSg
'house'	'common noun, sing, fem'
dwejra	NCFSgDim
'house-DIM'	'common noun, sing, fem, diminutive'

Degree

The comparative distinction in Maltese is explicitly marked, as shown in (14b), but the superlative is constructed by prefixing the definite article *(i)l-* before the comparative (14c).

(14) a. ktieb sabiħ
book nice
' a nice book' (positive)

b. ktieb isbaħ
book nice-COMP
'a nicer book' (comparative)

c. l- isbaħ ktieb
DEF- nice-COMP book

'the nicest book' (superlative)

While the *degree* attribute under the *adjective* category distinguishes between positive, comparative and superlative, the latter emerges as a composite of the definite article and the comparative. Tags are provided for the comparative, the positive being a default value for the adjective lacking comparative marking; the superlative can be accounted for via a lexical rule of the following type:

(15) $AT + AJCP \Rightarrow AJSP$

where AT is 'definite article', AJCP is 'comparative adjective' and AJSP is 'superalitve adjective'. Thus, at word level, the superlative is annotated as a composite of article and comparative, with the superlative emerging as a *meta-tag*, akin to the 'combined tags' used in the annotation system of the Brown Corpus.

2.3 Clitics, particles and case markers

Enclitic pronouns
A separate *attachment* attribute was included under the category *pronoun/determiner* in the tagset in order to account for a set of enclitic pronouns which attach to verbal, nominal or prepositional heads. Depending on their host, the pronouns have different case properties. Specifically, an enclitic pronoun attached to a verbal or prepositional head assigns accusative case, expressing the direct object; attachment to a nominal head assigns genitive case, expressing the possessor (16). Once again, the enclitic pronouns are distinguished from their host in annotation using the + sign.

(16)

Word	Tag
lil*u* 'to him'	API + PEN3SgMAc 'independent preposition' + 'enclitic pronoun, 3rd sing, masc, accusative'
lil*ha* 'to her'	API+ PEN3SgFAc 'independent preposition' + 'enclitic pronoun, 3rd sing, fem, accusative'
qatl*u* 'he killed him'	VV3SgMPf + PEN3SgMAc 'main verb etc' + 'enclitic pronoun, 3rd sing masc, accusative'
qatil*hom* 'he killed them'	VV3SgMPf + PEN3PlAc 'main verb etc' + 'enclitic pronoun, 3rd plu, accusative'
omm*u* 'his mother'	NCSgF + PEN3SgMGv 'common noun, etc' + 'encclitic pronoun, 3rd sing masc, genitive'
omm*na* 'our mother'	NCSgF + PEN1PlGv 'common noun, etc' + 'enclitic pronoun, 1st plural, genitive'

Since the sets of enclitic pronouns attaching to verbs, nouns and prepositions are isomorphic, and case distinctions depend solely on the category of the head, including case information may be costly in terms of processing required for accurate annotation. Nevertheless, case distinctions, even if not included at first pass during the annotation process, are relatively easy to account for (for instance, through the use of lexical rules).

Prepositions
An *attachment* attribute for prepositions distinguishes between independent and procliticised prepositions. Most prepositions procliticise to an adnominal head, exhibiting phonological dependence. This is also reflected in the orthography. An example is given in (17), outlining the distinction between independent and procliticised prepositions. Both are tagged independently from their host.

(17)

Example	Tag (preposition only)
fi djar (in houses)	API 'independent preposition'
f'dar (in house)	APPR 'procliticised preposition'

In addition, the EAGLES scheme provides a useful distinction under the attribute *type*, between *prepositions* and *fused preposition-article*. The latter is relevant for Maltese, as shown in (18).

(18)

Example	Tag (preposition only)
fid-dar (in-DEF-house)	APAT 'fused preposition-article'
mid-dar (from-DEF-house)	APAT 'fused preposition-article'

3. Sample annotation scheme for Maltese

The following table illustrates the complete information encoded in the tagset discussed in relation to the examples in §2. The tagset follows the general outline of the EAGLES recommendations, with each category having a set of attributes with a set of values for each attribute. The attribute-value structure is useful, since the re-interpretation of tags as feature structures is facilitated in this manner.
In the table, only the relevant attributes and values are shown, other elements found in the EAGLES standard not included in the tagset being omitted. As noted above, certain attributes were added to the original scheme, while the inclusion of other attributes was made in the *unique/unassigned* category, an expedient which was considered undesirable and which is currently the focus of attempts to revise the tagset.

Major Categor-ies	**i**	**ii**	**iii**	**Iv**	**V**	**vi**	**vii**	**viii**

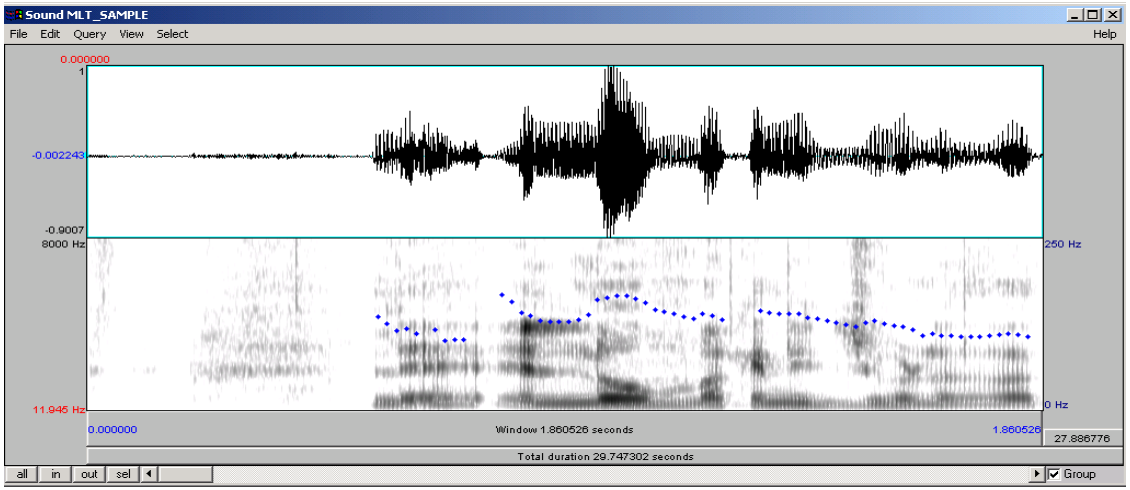
N noun	**Type** 1 Common 2 Proper	**Gender** 1 Masculine 2 Feminine	**Number** 1 Singular 2 Plural 3 dual 4 collec-tive	**Dimension** 1. diminutive				
V verb	**Person** 1 First 2 Second 3 Third	**Gender** 1 Masculine 2 Feminine	**Number** 1 Singular 2 Plural	**Mood** 3 Imperative	**Tense** 1 Past (kien) 2 Non-past	**Voice** 1 Active [=active participle] 2 Passive [=passive participle]	**Status** 1 Main 2 Auxiliary (kien) = tense marker	**Aspect** 1 perfective 2 imperfective
AJ adjective	**Degree** 1 Positive 2 Comparative 3 superlative	**Gender** 1 Masculine 2 Feminine	**Number** 1 Singular 2 Plural	**Dimension** diminutive				
PD pronoun /determiner	**Person** 1 First 2 Second 3 Third	**Gender** 1 Masculine 2 Feminine	**Number** 1 Singular 2 Plural	**Case** 1 Genitive 2 Dative 3 Accusative	**Attachment** 1 independent 2 enclitic			
AT article	**Article Type** 1 Definite							
AV adverb	**Degree** 1 Positive 2 Comparative 3 superlative							
AP adposition	**Type** 1 Preposition 2 Fused Prep-article	**Attachment** 1 indepent 2 proclitic						

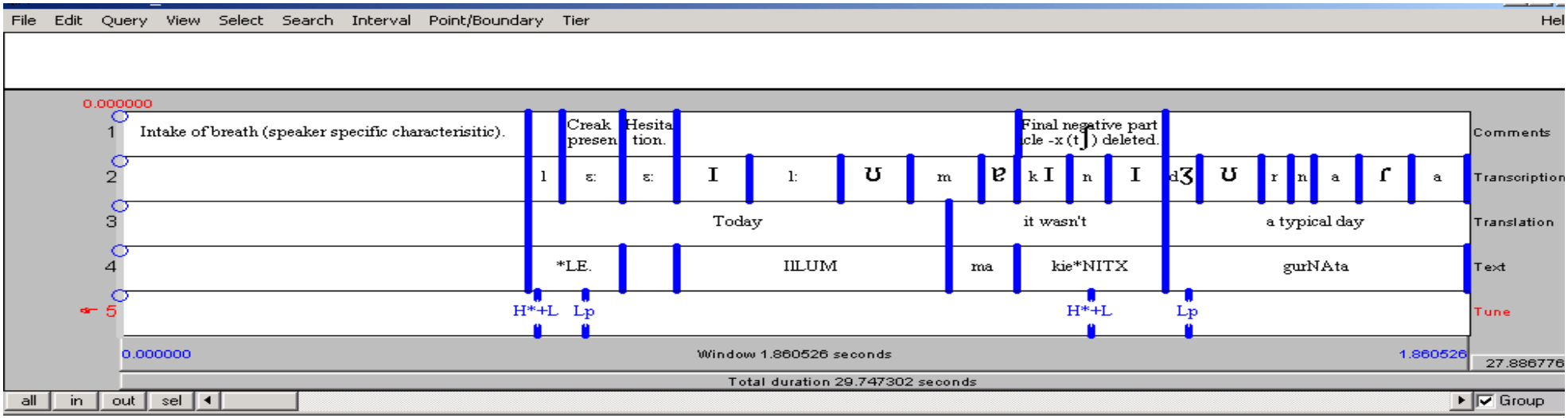
C conjunction	**Type** 1 Coordinating 2 Subordinating	**Coord Type** 1 Simple 2 Correlative						
NU Numeral	**Type** 1 Cardinal 2 Ordinal	**Gender** 1 Masculine 2 Feminine	**Number** 1 Singular 2 Plural					
I Interjection								
U unique / unassigned	**Type** 1 Multiword utterance 2 negator 3 aspect marker 4 dative case-assigning particle							
R Residual	**Type** 1 Foreign word 2 Formula 3 Symbol 4 Acronym 5 Abbreviation 6 Unclassified	**Number** 1 Singular 2 Plural	**Gender** 1 Masculine 2 Feminine 3 Neuter					
PU Punctuation	**Type** 1 Sent-final 2 Sent-medial 3 Left-parenthetical 4 Right-parenthetical							

Under the category *verb*, the active/passive values of the *voice* attribute distinction are specified as applying to the active and passive participles. This may clash with the schema that is valid for other languages. Hence, it should be understood as provisional. The possibility of there being a separate *participle* attribute encoding the active/passive distinction for Maltese is not excluded.

4. Annotation of spoken discourse

To date, no annotation scheme has been agreed upon for Maltese transcribed spoken discourse. The present section therefore limits itself to exemplifying the process of annotation of oral discourse, through the following two examples. The sample provided in the next two figures is taken from a studio recording of a conversation between two males who were aware that they were being recorded. Labelling and segmenting was carried out using PRAAT version 4.0.


Sound MLT_SAMPLE
File Edit Query View Select
Help
0.000000
1
-0.002243
-0.9007
8000 Hz
250 Hz
0 Hz
11.945 Hz
0.000000
Window 1.860526 seconds
1.860526
27.886776
Total duration 29.747302 seconds
all
in
out
sel
Group


File Edit Query View Select Search Interval Point/Boundary Tier
Hel
0.000000
1
Intake of breath (speaker specific characterisitic).
Creak presen
Hesita tion.
Final negative part icle -x (tʃ) deleted.
Comments
2
l ɛː ɛː ɪ lː ʊ m ɐ k ɪ n ɪ dʒ ʊ r n a ɾ a
Transcription
3
Today
it wasn't
a typical day
Translation
4
*LE.
lILUM
ma
kie*NITX
gurNAta
Text
5
H*+L
Lp
H*+L
Lp
Tune
0.000000
Window 1.860526 seconds
1.860526
27.886776
Total duration 29.747302 seconds
all
in
out
sel
Group

In the above figures, five annotation tiers have been used.

The Comments Tier

This tier includes information about phonetic features which were either idiosyncratic, such as intake of breath and creaky voice, or more systematic, for example aspiration and deletion. Some of this information would probably have been better included as part of a multi-layered transcription involving different degrees of broadness/narrowness.

The Transcription Tier

The transcription given on this tier includes information about certain co-articulatory effects which would probably have been better dealt with elsewhere. Some examples include the transcription of a single /m/ at the boundary between the /m/ at the end of *illum* and the /m/ at the beginning of *ma*, the transcription of /£/ but no final /¢/ at the end of *ma kienitx* where the latter is deleted in the context of the following /£/ and the transcription of a tap /F/ in *ġurnata* to show that the speaker substitutes the /t/ by a tap. Although there is as yet no standardised convention for the annotation of speech at the phonetic level, the symbols used are based on Borg and Azzopardi-Alexander (1997).

The Translation Tier

This tier provides an English translation of the Maltese text. It is intended to provide a guide to the more "global" meanings of larger units. A gloss may be a necessary extra.

The Text Tier

Although the text given on this tier is given using standard Maltese orthography it has not been possible to represent the Maltese graphemes ċ, ġ, ż, ħ, għ. A standard way for transcribing fillers, of which there is one example in the sample provided, *Ee* 'Eh', is necessary. Moreover, a deviation from standard orthography has been made in order to show how the tonal material transcribed on the tonal tier is associated with the text: stressed syllables are capitalised wherever they occur in the text while stressed syllables bearing the main stress of the phrase are indicated by means of an asterisk.

The Tune Tier

A tonal analysis based on an Autosegmental-Metrical framework is provided on the final tier. The analysis at this level follows work by Vella (1995 and 2003).

References

Gatt, A. (2001). An XCES-EAGLES compatible tagset for Maltese. Technical report, MaltiLex Project, University of Malta.

Borg, A., and Azzopardi-Alexander, M. (1997). *Maltese: Descriptive Grammar*. London: Routlege.

Vella, A. (1995). *Prosodic Structure and Intonation in Maltese an dits Influence on Maltese English.* Unpublished PhD Thesis, University of Edinburgh.

Vella, A. (2003). Phrase accents in Maltese: Distribution and realization.*Proceedings of ICPhS 2003,* Barcellona, Spain.