Construction and Annotation of a Maltese Corpus

Joe Caruana
joec@ling.ed.ac.uk

ABSTRACT

This paper describes work in progress towards the creation of a morpho-syntactically annotated computer corpus of Maltese. Hitherto, the language has not been the object of much research in either corpus-linguistics or natural language processing, and is largely bereft of the apparatus of computer-linguistic resources and tools available for work on other languages. The texts making up the existing corpus data are preponderantly drawn from written sources, the large majority being newspaper writing. A small proportion of the texts is drawn from transcribed spoken data, largely of broadcast nature. The tension between the availabilty of texts and the linguistic desiderata of "representativeness" and "balance" in corpus design is briefly discussed. The paper then proceeds to outline the choices made with regard to encoding the corpus data. The adoption of XCES/EAGLES corpus architecture and markup guidelines is briefly discussed and motivated. The adaptations that certain linguistic features of Maltese necessitate in both EAGLES and XCES standards are described, and illustrated with reference to the number system of Maltese nouns, the root and pattern morphology of Semitic Maltese, and cliticisation. The paper concludes with a prospectus of further work on the project.

1. INTRODUCTION

The collection of written Maltese texts for inclusion in a first electronic corpus of the language started in 1998, within the MALTILEX Project (Rosner at al. 1998). The work was motivated by a recognition of the contribution that a corpus could make both to the linguistic study of Maltese and to the growth of computer linguistic work on the language, as well as the development of resources applicable in teaching and lexicography. The central preoccupation at the initial stages of the project was the creation of a first machine-readable lexicon of Maltese, a foundational resource for any NLP work on the language.

Work on the corpus, both in terms of its collection and its markup with linguistic information, is still very much in progress. In reporting on the current state of the project, this paper will first give a brief description of the corpus material, and then outline the markup strategies that have been adopted, together with their motivation. The immediate prospectus for further work concludes the paper.

2. CORPUS TEXTS AND ENCODING

A number of challenges had to be overcome in collecting the corpus data. Lack of manpower and funds meant that it was not feasible to have corpus data input by hand. Similarly, the absence of a machine-readable lexicon of Maltese meant that scanning in printed texts was not feasible, as OCR software would not be able to progress beyond basic character recognition to actual word recognition. These two factors meant that only texts that were already in electronic format could be considered for inclusion in the corpus. Given these constraints, it should not be surprising that the collection of texts has been largely opportunistic in nature. In an ideal world, it would have been

preferable to proceed on the basis of a predefined taxonomy of text types and carefully calibrate the relative weightings of different genres and registers within the corpus according to a principled notion of their balance and relative weightings, with a view to achieving the holy grail of linguistic representativeness. While every attempt has been (indeed, still is being) made to diversify and expand the content of the corpus, the immediate thrust of the present approach has been geared more towards creating a collection of texts that could serve as a good test bed for developing corpus processing tools and procedures. Revising the structure of the corpus so that it can make better-justified claims to being representative of modern Maltese is a task deferred to a future date.

2.1. Composition of the corpus

At present, the corpus contains some 2 million words of text. None of the corpus material is older than 1990. Most of the corpus contents are complete texts, but there is a significant proportion of excerpts taken from longer published works. Corpus composition is broken down in Figure1 below. The large bulk of the corpus data is drawn from the written language, only a tenth being transcribed spoken data. Within the written component of the corpus, newspaper texts predominate, the main subject areas being domestic political news and commentary, international current affairs, media, economy and business. The current representation of newspaper text within the corpus is judged sufficient, and no more are being collected. In contrast, the part of the corpus drawn from book publishing is in need of expansion, both as regards fiction and non-fiction titles. Obtaining material and copyright permission from book publishers has proved a much more uphill task than was the case with newspapers. However, one very positive development is an agreement with the publishers of an important and extensive non-fiction series. Texts from this series, conceived as a millennium project encyclopaedia of all things Maltese, are especially useful because of their wide domain coverage. Ephemera (personal letters, handbills, etc.) are underrepresented at present, as are texts of an administrative or legal nature.

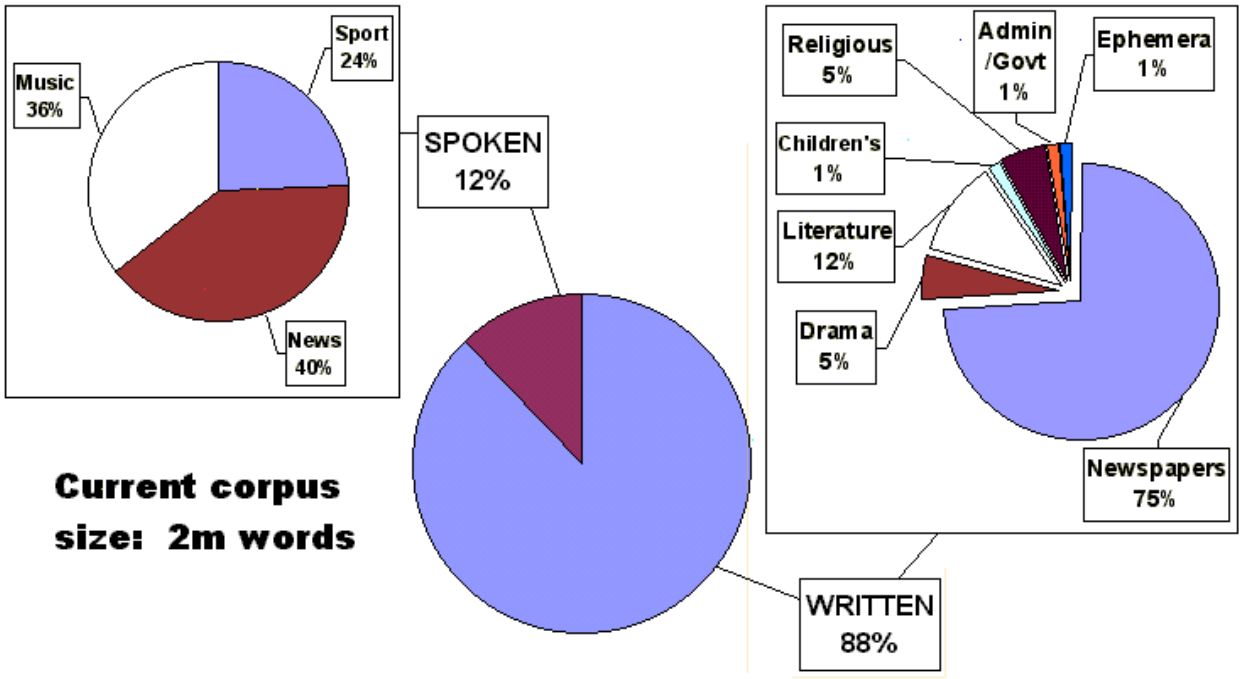

Figure 1: Current composition of the Corpus

The spoken data is made up of transcriptions of broadcast (radio) speech, from three main areas: music, sport and news broadcasts. The music (and to a somewhat lesser extent, sport)

programmes contain a great proportion of interactions that approximate spontaneous conversational speech, while the news programmes are texts that are exclusively of a "written-to-be-spoken" nature. Other data is currently being added to the spoken part of the corpus. In particular, there is a set of interviews graciously contributed by Martine Vanhove, in which speakers of a dialectal variety of Maltese interact and recount narratives. This is especially interesting in that the rest of the data is very much within the standard variety of the language.

2.2. Encoding

The writing system of Maltese, codified as recently as the nineteen-twenties, is based on the Latin alphabet, but includes the additional characters •, •, •, •, ie and g•. The last two are digraphs, but, for some reason, they have not been incorporated as such in the Unicode tables. Furthermore, they are never treated as units by Maltese typists and typesetters, being composed at the keyboard from the characters 'g' and '•' (or more often than not plain ASCII h, see below), 'i' and 'e' respectively.

	ISO10646-1 ISO8859-3	ASCII representation		ISO10646-1 ISO8859-3	ASCII representation
•	010B	_c	•	010A	_C
•	0121	_g	•	0120	_G
•	0127	_h	•	0126	_H
•	017C	_z	•	017B	_Z
ie	--	_i	Ie / IE	--	_I
g•	--	_y	G• / G •	--	_Y

Table 1: Maltese-specific characters, character codes and basic ASCII representation

Agreed standards in the electronic encoding of Maltese and the definition of a keyboard for the language are only now receiving official attention. In the absence of such standards, Maltese writers long used to typing on computer keyboards/systems that only support the ASCII character set, have for decades resorted to the strategy of substituting the Maltese-specific characters with those ASCII characters they most resemble. This works well enough in practice, given that adult Maltese readers are readily able to distinguish between occurrences of either, even where the word is devoid of context, as in those in line (1b) below:

(1)	a.	kabo••a	ga••a	•ieg•el	•uha	•militizzazzjoni
	b.	kabocca	gagga	gieghel	huha	zmilitizzazzjoni
		'cabbage'	'cage'	'he forced'	'her brother'	'demilitarisation'

However, from the point of view of corpus encoding, this raises some problems. The presence of Maltese language on the web is almost exclusively in the form of texts generated purely in ASCII. There is also a considerable body of texts of a legal and administrative nature that are written in this way, for example, the CD-ROM collection of the Laws of Malta and a number of judicial deliberations/sentences. In spite of their ready availability, therefore, their use within the corpus is not unproblematic, as considerable pre-processing is required to bring them to conformity with the writing system of the language.

The corpus exists in two different encodings. The most basic level, to which all documents obtained for inclusion in the corpus are first reduced, involves the use of a work-around whereby the Maltese-specific characters, including the two digraphs, are represented by means of an

underscore+letter combination. This is shown in the 'ASCII representation' columns in Table 1, and (2) below exemplifies this as applied to the words in (1a):

(2) kabo_c_ca ga_g_ga _*g*_i_yel _huha _zmilitizzazzjoni

This basic ASCII encoding guarantees maximum platform- and software-independence. In parallel, the corpus is maintained Unicode (ISO10646-1) encoding. Unicode brings together (in addition to many more) the characters contained within the ASCII and Latin Extended-A (ISO8859-3) character sets referred to in Table 1 above, as well as the Latin-1 Supplement (ISO8859-1), a character set which is of relevance in that it contains the grave-accented vowels that occur in many Romance loans in Maltese. This will become the default encoding as support for Unicode becomes more widespread.

3. STRUCTURE AND MARKUP

The corpus is marked up in Extensible Markup Language (XML), a development and subset of SGML that allows for the manipulation of data structures as well as textual content. The choice of this markup framework was motivated by a variety of considerations, foremost among which was the desire to situate the present work of corpus building and annotation as closely as possible within similar work being carried out in a number of projects and languages. While it is still a developing framework, XML and its related technologies have attracted great attention from across the computing world, and are at the heart of an increasing number of corpus-linguistic and NLP tools and projects. The adoption of XML should therefore allow for increased ease of data interchange, comparison and reuse, as well as interoperability of corpus-processing tools produced at different sites.

3.1. Standards and recommendations

In line with this recognition of the importance of interchange and standardisation, it was decided to mark up the corpus in conformity with a set of recommendations enjoying wide acceptance within the corpus linguistic community. There have been a number of such attempts at standardising approaches and frameworks for the collection and annotation of linguistic corpora and related data structures. Within the framework of Standardised General Markup Language (SGML), the Text Encoding Initiative (Burnard et al. 1995), a project aimed at setting standards and recommending best practice in humanities computing, devoted attention to the area of corpus linguistics. The Corpus Encoding Standards (CES) group further developed and refined the TEI recommendations, and produced an integrated set of guidelines, recommendations and Document Type Definitions (DTDs) for language corpora annotated at different levels of linguistic analysis (Ide 1998). The Expert Advisory Group on Language Engineering Standards (EAGLES) has also been very influential in the area of data-base and language software development, especially within the European Union (and by extension Europe beyond the EU). The latest development in this interconnected chain of initiatives is the publication of the XCES (Ide et al. 2000) which essentially ports the CES guidelines/recommendations from SGML to XML.

3.2 Adapting the EAGLES guidelines

The EAGLES guidelines for morphosyntactic tagging of language corpora (Leech and Wilson 1996) propose a common framework that tries to identify those sets of concepts and grammatical entities that are of relevance to the analysis and description of European languages. They propose a common set of minimum markup in terms of grammatical category, setting up thirteen obligatory categories: N noun; V verb; AJ adjective; PD pronoun/determiner; AT article; AV adverb; AP

adposition; C conjunction; NU numeral; R residual; and PU punctuation. Each category carries a set of attributes, and each of these in turn have sets of possible values. For each attribute, only one value can be paired with a given attribute at any one time, such that it is possible to construct a full and unequivocal morphosyntactic descriptor for a corpus token in terms of Category and Attribute+Value pair. Table 2 below lays out the matrix for the Noun category.

i	Category	**N** Noun					
ii	Type	**1** Common	**2** Proper				
iii	Gender	**1** Masculine	**2** Feminine	**3** Neuter	***4*** *Common (Danish, Dutch)*		
iv	Number	**1** Singular	**2** Plural				
v	Case	**1** Nominat-ive	**2** Genitive	**3** Dative	**4** Accusat-ive	***5*** *Vocative*	***6*** *Indeclinable (Greek)*
vi	Countability	**1** Countable	**2** Mass				

Table 2: EAGLES Noun morphosyntactic descriptor codes (Leech & Wilson 1996)

The Roman numerals in Table 2 indicate each successive item's position in a left-to-right sequence. Thus, the EAGLES-compliant code for the morphosyntactic description of the English noun 'book' would be **N10101**, parsed from left to right to read: Category 'Noun'; Type 'Common' ; Gender 'Not applicable'; Number 'Singular'; Case 'Not applicable'; Countability 'Countable'.

3.2.1 Number oppositions in Maltese nouns

This EAGLES descriptor code system is not, as it stands, immediately applicable to Maltese, as the oppositions it presumes are not exhaustive of those obtaining in the language. For example, the number system of Maltese nouns is more complex than a simple singular/plural opposition, adequate though it is for the majority of nouns (3)[1], but also adds Dual number (4) for some nouns

(3) a. *ktieb* 'book'; *kotba* 'books'
b. *e•empju* 'example'; *e•empji* 'examples'
c. *si••u* 'chair'; *si••ijiet* 'chairs'

(4) *sieg•a* '(1) hour'; *sag•tejn* '(2) hours'; *(3+) sig•at* '(3+) hours'

A large class of nouns have a different three-way number opposition between Singular, Collective and what is variously called a Singulative or Determinate Plural for number range two to ten (note that n=11-19 take the form *n-il larin•a*; n=20+ take the form *n larin•a*)

[1] Note that (3a,b,c) exemplify three distinct patterns of pluralisation in Maltese

(5) *larin•a* '(1) orange'; *larin•* 'oranges' (as class); *(2-10) larin•iet* '(2-10) oranges'

Finally, yet another distinction comes into play with some nouns, which add what has been called the Indeterminate Plural form

(6) *dubbiena* '(1) fly'; *dubbien* 'flies' (as class); *(2-10) dubbiniet* '(2-10) flies'; *dbieben* '(large but indeterminate nr. of) flies'

Therefore, row (iv) Number in Table 2 above has to be amplified if we are to account for the number system of Maltese nouns. Table 3, which shows only those matrix cells that are applicable to the description of Maltese nouns, shows these additional values for the Number attribute, together with the addition of Verbal Noun in row (ii)

i	Category	**N** Noun					
ii	Type	**1** Common	**2** Proper	**3** Verbal			
iii	Gender	**1** Masculine	**2** Feminine				
iv	Number	**1** Singular	**2** Plural	**3** Collective	**4** Determin-ate	**5** Indeter-minate	**6** Dual

Table 3: Maltese-specific modifications to EAGLES morphosyntactic descriptor codes for Noun category

Note that the matrix rows for Case and Countability, which in Table 2 occupied positions (v) and (vi) in the descriptor code matrix, have been dropped. Case is not applicable to Maltese, and the Countable/Mass distinction is redundant, given a Number system as described above: 1 Singular, 2 Plural, 4 Determinate Plural, and 6 Dual are Countable, while 3 Collective and 5 Indeterminate Plural are Mass. Listing (7) below shows the application of these descriptor codes to some Maltese nouns

(7)

si••u 'chair'	N111	*dbieben* '(indeterminate nr. of) flies'	N105
si••ijiet 'chairs'	N102	*sag•tejn* 'two hours'	N106
dubbiena 'fly'	N121	*Sandra*	N221
dubbien 'flies (class)'	N113	*mixi* 'walking'	N313
dubbiniet '(2-10) flies'	N104	*mixjiet* 'walks (acts of walking)'	N304

Similarly to the adaptations outlined above with respect to the number system of nouns, other changes have been made to the descriptor matrices for other categories. The result is a system of morphosyntactic descriptor codes that can be supplied in the <msd> element of the XCES-compliant markup. These descriptor codes are readily mappable onto any corpus-specific tags that may be adopted, and which are carried by <ctag> elements.

3.3 Adaptation of XCES

At the level of the individual corpus token and its morphosyntactic annotation, the data structure proposed by XCES does not depart significantly from that put forward earlier by the CES. This

structure allows for a token element <tok> to comprise the orthographic representation of the token/word as it appears in the text <orth>, one or more lexical analyses of the token <lex>, each associated with its lemma or base form <base>, the tag used in the corpus to represent the morpshosyntactic analysis <ctag> and its representation in the EAGLES format <msd>. The element <disamb> encapsulates the human- or machine-disambiguated morphosyntactic analysis.

3.3.1 Roots and patterns

The mixed nature of Maltese is reflected in its morphology, which has traits from both Romance stem and affix morphology and a Semitic root and pattern morphology. Such a language was obviously not taken into account by the compilers of the EAGLES recommendations for morphosyntactic annotation, looking as they were to the languages of the European Union countries and those of the Eastern European nations beyond.

From the point of view of corpus building and exploitation, the importance of annotating Maltese words with their consonantal root, where appropriate, lies not in any concern with the diachronic. Synchronically, a shared root allows us to identify a common semantic content, operated upon and changed by means of regularly varying patterns of affixation, each of which tends to be associated with a particular behaviour or meaning. It is important to explicitate and retain this information within the corpus, with an eye to future work in information extraction and discourse mapping within the corpus texts.

Table 4 below (adapted from Mifsud 1995: 36) illustrates this feature of the language with respect to verb patterns, called *forom* 'forms' in Maltese: nine trilateral patterns (I-X; pattern IV is no longer represented in the lexicon), and two quadriliteral patterns (QI-II). The pre- and infixes are shown in bold, while the rightmost column presents the relative codes used in the corpus annotation (cf. 'pattern' attribute below).

I	` 1 v 2v 3	Basic active meaning, transitive and intransitive	T1
II	` 1 v **2** v 3	Intensive of I transitive; transitive of I intransitive	T2
III	` 1• 2 v 3	Transitive of I intransitive	T3
V	**`t** 1 v 22 v 3	Passive and/or reflexive of II	T5
VI	**`t** 1 • 2 v 3	Passive and/or reflexive of III	T6
VIIa	**`n** 1 v 2 v 3		T7A
b	**`nt** 1 v 2 v 3	Passive and/or reflexive of I	T7B
c	**`n** 1 **t** v 2 v 3		T7C
VIII	` 1**t** v 2 v 3	Reflexive (and/or passive) of I	T8
IX	` 1 2• 3	Inchoative, acquisition of quality (colour)	T9
Xa	**`st** v1 v 2 v 3	(Diachronic) inchoative	T0A
b	**`st** 1 v 22 v 3		T0B
QI	` 1 v 23v 4	Basic active meaning (transitive and intransitive)	Q1
QII	**`t** 1 v 2 3v 4	Passive and/or reflexive of QI	Q2

Table 4: Root & Pattern permutations of Semitic Verbs in Maltese

These consonantal roots are a feature of words in categories other than verbs. Listing (8) below shows nouns having KTB as their root: (8a-h) are in common use, while (8i-j) are just two of the archaic lexical items reported by Aquilina (1987) in the entry for root KTB. These, though not used in current standard Maltese, are still readily interpretable by speakers of the language.

(8) a. *ktieb* ‘book’
Noun.Common.Masculine.Singular
b. *kotba* ‘books’
Noun.Common.Masculine.Plural
c. *ktejjeb* ‘booklet’
Noun.Common.Diminutive.Masculine.Singular
d. *kittieb* ‘writer’
Noun.Common.Agentive.Masculine.Singular
e. *kittieba* ‘female writer’ | ‘writers’
Noun.Common.Agentive.Feminine.Singular | Noun.Common. Agentive.Plural
f. *kittibin* ‘scribes / writers’
Noun.Common. Agentive.Plural
g. *kitba* ‘item / act of writing’
Noun.Common.Verbal.Feminine.Singular
h. *kitbiet* ‘items /acts of writing’
Noun.Common.Verbal.Plural
i. *ktib* ‘act of writing’
Noun.Common.Verbal.Feminine.Singular
j. *mikteb* ‘desk (place of writing)’
Noun.Common.Masculine.Singular

The XCES structure has been minimally modified to allow for the annotation of the Maltese corpus tokens with root information, where appropriate. The modification takes the form of the inclusion of an optional ‘root’ attribute carried by the <base> element, where ‘root’ is given as an uppercase three- or four-consonant series. In addition, where appropriate, another optional attribute ‘pattern’ is carried by the same element. The codes for this appear in the rightmost column of Table 2. Note that the <base> element is a node within both <lex> and <flex>, as shown in listings (13) and (14) below.

3.3.2 Clitics

The XCES <chunk> element allows for the treatment of lexical items that contain whitespace as single (multi-word) units, e.g. *kelb il-ba•ar* ‘shark’, which can be recognised as such at the tokenisation stage. However, no provision is made for compositional elements occurring below the <lex> level. This is a difficulty, as Maltese word-forms can be rich in clitic elements. For example, most verbs can carry up to four suffixal elements, as shown in (9b-h) below:

(9) a. Kiteb ismu (*kiteb* V)
‘He wrote his name’
b. Ma kitibx ismu (*kitib + x* V+Neg)
‘He did not write his name’
c. Kitibni fir-re•istru (*kitib + ni* V+Pro[dO])
‘He entered my name in the register’
d. Ma kitibnix fir-re•istru (*kitib + ni + x* V+Pro[dO]+Neg)
‘He did not enter my name in the register’
e. Kitibli ittra (*kitib + l + i* V+Prep +Pro[iO])
‘He wrote a letter to/for me’
f. Ma kitiblix ittra (*kitib + l + i + x* V+Prep +Pro[iO]+Neg)
‘He did not write a letter to/for me’

g. Kitibhieli bil-lapes (*kitib + hie + l + i* V+Pro[dO]+Prep+Pro[iO])
'He wrote it to/for me in pencil'

h. Ma kitibhilix bil-lapes (*kitib + hi + l + i + x* V+Pro[dO]+Prep+Pro[iO]+Neg)
'He did not write it to/for me in pencil'

This propensity for adding enclitics has the potential to dramatically increase the total number of morphosyntactic tags used in the corpus. This is especially true in the case of verbs. The starting position looks very promising from the point of view of someone wishing to carry out automatic statistical tagging of Maltese texts. The verb system encodes morphological distinctions between two moods alone, the Indicative and the Imperative, and only two tenses, the Imperfect and the Perfect. Other moods and tense/aspect combinations are formed periphrastically. Person, gender and number of the verb subject together account for a seven-way distinction in each of the Perfect and Imperfect Indicative, and only four distinct forms in the Imperative. This is a very promising starting-off point for anyone wanting to carry out stochastic morphosyntactic tagging on texts in the language, as a small tagset makes for greatest economy, in that the required manually-tagged training corpus need only be of modest size. The situation is rendered appreciably more fraught when these verb+clitic combinations are taken into account, as they would generate 706 separate tags were each verb-form and clitic(s) combination be tagged as a unit with a unitary morphosyntactic tag. Even were one to adopt the procedure proposed in Tufi• (2000), i.e. using hidden reduced tagsets to build competing language models, which can then be automatically reconciled/eliminated in successive passes, would still require an unreasonably large training corpus, given the constraints of the present project. The solution is to do more work in pre-processing, at the tokenisation stage, breaking up these complex lexical items into atomic tokens. This keeps the number of distinct tags down to a far more manageable 26. The tokeniser breaks up the three word sentence (10) into seven separate tokens, returning the listing in (11). Note that the `^(`caret) character (chosen in preference to the more intuitive plus sign because, unlike the latter, it does not occur independently in the corpus texts) indicates that the token it initiates is a fragment, and belongs in sequence with the token preceding it. The hyphen separating an article from the noun it determines is treated as part of the article token itself, rather than as a punctuation mark, which are tokens in their own right.

(10) *Kitibhieli bil-lapes.*
'He wrote it to/for me in pencil'

(11)
```
kitib   ^hie   ^l   ^i   bi   ^l-   lapes   .
```

The tokens can then be separately tagged for morphosyntax. The output on (11), in the form of plain text, whitespace-delimited linear strings, is shown in (12). Note that a plus sign at the start of a tag performs the same function with respect to the tag as the caret sign with respect to the token.

(12)
```
kitib     V.ID.PF.P3.MS.SG.
^hie      +PD.P3.FM.SG.
^l        +P.LL.
^i        +PD.P1.SG.
bi        P.
^l-       +AT.
lapes     N.
.         PU.FS.
```

In order to better account for the incorporation of word constituents into the system proposed by XCES and EAGLES, an element below the level of <lex> has been added to the relevant DTD. This element, <flex> for 'lexical fragment', encapsulates each component of the composite word form, whether stem or affix, together with structural elements mirroring those pertaining to the larger unit. Thus, each <flex>, carrying an identifier attribute, contains an <orth> element showing the orthographic form of each fragment within the full wordform, a <base> element, optionally carrying a root attribute to show the consonant root of the lexical item and a pattern attribute to indicate verb pattern (3.3.1 above), a <msd> element, shown here only for the noun •*ars* in (14), as other categories were not discussed in section 3.2 above, giving the EAGLES code for the fragment descriptor, and a <ctag> giving the same information in the corpus-specific code. In this system, the word *jiktibhielkom* 'he writes it to/for you (pl.)' is represented as follows:

(13)
```
<tok id='1'>
  <orth>jiktibh_ilkom</orth>
  <lex id='1.1'>
     <base root='KTB' pattern='T1'>kiteb</base>
     <flex id='1.1.1'>
        <orth>jiktib</orth>
        <base>kiteb</base>
        <msd></msd>
        <ctag>V.ID.IP.3P.MS.SG.</ctag>
     </flex>
     <flex id='1.1.2'>
        <orth>^h_i</orth>
        <base>hija</base>
        <msd></msd>
        <ctag>+PD.PR.3P.SG.</ctag>
     </flex>
     <flex id='1.1.3'>
        <orth>^l</orth>
        <base>lil</base>
        <msd></msd>
        <ctag>+PP.LL.</ctag>
     </flex>
     <flex id='1.1.4'>
        <orth>^kom</orth>
        <base>intom</base>
        <msd></msd>
        <ctag>+PD.PR.3P.PL.</ctag>
     </flex>
     <msd></msd>
     <ctag>
        V.ID.IP.3P.MS.SG.+PD.PR.3P.SG.+PP.LL.+PD.PR.3P.PL.
     </ctag>
  </lex>
  <disamb>
     V.ID.IP.3P.MS.SG.+PD.PR.3P.SG.+PP.LL.+PD.PR.3P.PL.
  </disamb>
</tok>
```

The use of the + symbol within the <msd>, <ctag> and <disamb> elements indicates the compositional character of the information, sequential information being given by the relative position. The symbol concatenates the morphosyntactic description on its immediate right to the one immediately preceding it. This symbol has been introduced to make possible a linear string representation of the information given within successive <flex> nodes, in effect giving a unitary morphosyntactic tag to the wordform as a whole, without incurring the penalties that a ballooning tagset would impose on corpus processing.

The representation of an ambiguous token carries more than one <lex> element, the preferred reading being identified within a <disamb> element after either manual or automatic disambiguation. This structure can be exemplified by the word •*arsu*, which is variously interpretable as a possessive (so-called 'construct state') noun and pronoun composite 'his looking', a second person plural imperative verb 'look', and a third person plural indicative perfect verb 'they looked' below (note that the <disamb> element is not filled in this example, as the identification of its value would depend on context):

(14)
```
<tok id='1'>
  <orth>_harsu</orth>
  <lex id='1.1'>
     <base root='_HRS' pattern='T1'>_hares</base>
     <msd></msd>
     <ctag>V.IM.2P.PL.</ctag>
  </lex>
  <lex id='1.2'>
     <base root='_HRS' ptrn='T1'>_hares</base>
     <msd></msd>
     <ctag>V.ID.PF.3P.PL.</ctag>
  </lex>
  <lex id='1.3'>
     <base root='_HRS'>_hars</base>
     <flex id='1.3.1'>
        <orth>_hars</orth>
        <base root='_HRS'>_hars</base>
        <msd>N313</msd>
        <ctag>N.VN.MS.CL.</ctag>
     </flex>
     <flex id='1.3.2'>
        <orth>^u</orth>
        <base>huwa</base>
        <msd></msd>
        <ctag>+PD.PR.3P.MS.SG.</ctag>
     </flex>
     <msd></msd>
     <ctag>N.VN.MS.CL.+PD.PR.3P.MS.SG.</ctag>
  </lex>
  <disamb></disamb>
</tok>
```

4. FURTHER WORK

As was noted at the outset of this paper, this is very much an ongoing project. The collection of corpus texts will continue, and the proportions and relative weightings of the registers and domains will be revised in order to make the corpus as representative as possible of contemporary standard Maltese, at least within the written medium. Further work on developing a better lemmatiser is also required. Tagging the corpus data with morphosyntactic information is the main focus at the moment: a manually-tagged training corpus of some forty thousand words of text is close to completion, and, once this has been validated by other human annotators, experiments will begin to apply a variety of existing statistical taggers, for example Thorsten Brants' TNT and Oliver Mason's QTAG, to the task of annotating the corpus data as a whole.

References

Aquilina, J. 1987. *Maltese English Dictionary*. Malta: Midsea Books

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

Ide, N. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*. Granada: LREC. pp.463-70.

Ide, N., P. Bonhomme, and L. Romary 2000. XCES: An XML-based Standard for Linguistic Corpora. *Proceedings of the Second Language Resources and Evaluation Conference*. Athens: LREC. pp.825-30.

Leech, G., and A. Wilson. 1996. EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. (EAGLES Document EAG-TCWG-MAC/R). Pisa: Istituto di Linguistica Computazionale

Mifsud, M. 1995. *Loan Verbs in Maltese: a descriptive and comparative study*. (Studies in Semitic Languages and Linguistics). Leiden: E.J.Brill

Rosner, M. (ed.) 1998. *Computational Approaches to Semitic Languages: proceedings of the workshop*. Montreal: COLING-ACL

Rosner, M., J. Caruana, and R. Fabri. 1998. Maltilex: a computational lexicon for Maltese. In Rosner (ed.) 1998. pp. 97-101

Tufi•, D. 2000. Using a Large Set of EAGLES-compliant Morpho-syntactic Descriptors as a Tagset for Probabilistic Taggers. *Proceedings of the Second Language Resources and Evaluation Conference*. Athens: LREC. pp. 1105-1112