RULE BASED LEXICAL ANALYSIS OF MALTESE

Paul Micallef , Department of Communication and Computer Engineering, University of Malta
E-mail pjmica@eng.um.edu.mt

ABSTRACT

Since no computer based dictionaries exist for Maltese, the only analysis that can be made at present is rule based. The paper describes a rule based system taking into account the mixed origins, (semitic and romance), of Maltese to categorise function and verb words. The system separates the database, the rule formalisms and the rule definitions, enabling easier analysis and quicker changes to rules when necessary.

INTRODUCTION

Maltese is written in Roman script using 30 letters made up of six vowels and twenty four consonants. Of particular interest are two graphemes għ and ie which though written as two letters are actually considered as one grapheme in the language. A recent survey, based on the most well known Maltese dictionary [1], found that the origin of the words is 40% Semitic, 40% Romance (Italian) and 20% English. The linguistic processing has therefore to take into account this mixed structure. The purpose of this work was to analyse the words to try and distinguish function words and verb words in the running text, within a text-to-speech synthesis application. This is necessary to obtain pause information at appropriate word boundaries.
Therefore only a partial lexical analysis is being considered. The method used here involves a rule based system with separation between the rule definitions, the rule operations and the input data formalism. Two databases have the function words and the verb words respectively, to compare with the word under test. The rule formalisms are kept in small local tables pertaining to the various rules under test. In this way addition/deletion of the data is completely independent of the rule formalism, and rules can be added or amended independently of each other.

MALTESE LINGUISTICS

As in other Semitic languages, Maltese words of semitic origin do not have a stem to which affixes are connected, but rather use transfixes. The stem or root is made up of a number of consonants, which can never occur in isolation, and whose order cannot be altered. Transfixes are then added to the root, sometimes also with prefixes and suffixes. Transfixes are made up of a number of vowels and may include operation on consonants such as doubling the middle consonant, (geminate). On the other hand, words of Romance origin follow the usual pattern of a stem and affixes of the inflectional and derivational type to form other lexemes. For example the semitic derivations from k,t,b are

kiteb	to write
kitba	writing
ktib	writings
ktieb	book
ktejjeb	small book
kittieb	writer
mikteb	writing desk
nkiteb	to be written

while romance derivations from

ċert	certain
aċċerta	to ascertain
ċertezza	certainty
ċertezzi	certainties
ċertament	certainly
inċert	uncertain
ċertissmu	very sure

FUNCTION WORDS

Function words in Maltese are classified as *partiċelli* and *pronomi* and they are made up of pronouns, prepositions, conjunctions, interjections and adverbs. These can in turn be

distinct words	*hawn, hekk, għal, qabel, jien*
have personal pronoun suffixes	*għalina, qablek*
composite	*hawnhekk, għalhekk*
short phrases	*fuq il-qalb, sewwa sew*

In addition the definite article is added to the list of function words. The dictionary by Aquilina was used to obtain a comprehensive list of function words, for the database.
Some function words use inflectional morphemes as suffixes. The definite article and function words that assimilate the definite article change their final letter for some consonants, distinguished as *xemxin*. Therefore morphological analysis is essential to keep the function word database to a reasonable size. The function word types are also distinguished for the purposes of the application as

article	*il-, ir-*
pronoun	*jien, huma*
adverb	*hekk, sewwa*
conjunction	*u, jekk, mela*
prepositions assimilating the article	*mas-, fil-, ta'*

This subdivision is important as it helps in the syntactic analysis. The definite article is associated with a noun or adjective, the pronoun is associated with a noun phrase, a conjunction introduces a phrase, and the adverb indicates a verb phrase. Many function words can be of more than one type, and depend on the sentence syntax for the outcome.

VERB WORDS

Verb words in Maltese can be, like in other languages, conjugated in the present and past tense. There is no formal future tense, as in Italian, and auxiliary verbs are used to obtain other tenses. In general the present tense has prefixes to distinguish person, and suffixes to distinguish quantity. The past tense has suffixes to distinguish both person and quantity. In all of these the surface form of the stem can change. Like other languages there are numerous exceptions. Other verb words include the imperative, the negation of the verb and the passive and reflexive forms. Additionally the Maltese language tends to use suffixes extensively within verb words, to obtain generic accusative and dative object. For example

nikteb	I am writing
niktibhom	I am writing them
niktibhomlok	I am writing them, to you
kitbithieli	She wrote it (fem.) to me

The number of pronoun suffixes that can be added to every verb is considerable, [2].

LINGUISTIC ANALYSIS

To keep the same formalism, the same data structure is defined for the analysis of the function words and the verb words. This consists essentially of the stem consonants, the stem word, the part of speech, and the lexical group. The lexical group relates to the VC, (vowel consonant) sequence within the word. This order gives rise to different manipulations of prefixes and suffixes with the stem, and therefore different sets of morphological rules.

Function Word Analysis

The word under test is first checked for a valid function word. The analysis starts by looking for valid suffixes. The remaining stem is analysed for the consonants within the stem. The function word database is then examined for keys with the same consonants or more. Each corresponding entry has its lexical group which is then used for the test. The test is

according to the morphological rule appropriate for that lexical group. If a match is made the word is assigned the corresponding function word category stored in the function word database. The rule syntax is as follows. Removing suffix 1 results in a stem that has operations done on stem. Stem operations are denoted as

(Type)(Letter Strings)(Position) (1)

where Type can be

(+, , -) meaning addition, no addition, deletion

or (+ / -) meaning delete the left string and add the right string.

Letter Strings are ASCII strings to add or delete. In cases where no operation or only one type of operation is to be done, the rest of the field is left blank. Position is optional and denotes where in the stem the change should happen. The position is with respect to the end (right hand side of stem). Default is 1 and means abut to the stem. For example

(+)('e')(2) means add to the stem letter e at position 2 from end of the stem

(+/-)(a'/iegh) means delete the part iegh from the stem end and substitute with a'.

No operation is also valid. The decision for validity depends on whether the resulting stem, after stem operations, is a valid lexeme in the database.

Verb Word Analysis

If not a valid function word, a further test is made to check whether it is a valid verb word. Verb word analysis is initially different from function word analysis since verb words can also have prefixes. All possible consonant group sequences made up of two or more consonants from the surface word are considered, starting from the sequence with all the consonants in the surface word. Initially these groups are passed through the irregular verb list, then through the mute verb list, and then through the database. (Mute verbs are those that have a consonant in the stem that is missing in the surface form). Any consonant groupings found in the surface form that have entries in the dictionaries (irregular, mute or main) are potential candidates. For the first potential candidate the lexeme stem part defined as that part of the surface word that incorporates the lexeme consonants is isolated. If any prefix or suffix stems result, these are examined using the morphological rules for verbs. This results in either rejection since the affix stems cannot result in valid affixes, or valid affixes. If the affix stems have any remaining parts after valid suffix and prefix assignment, it is returned to the lexeme stem. The lexeme stem is now examined for a valid lexeme stem for the particular database lexeme under test. This is done by a separate rule set for each verb type. If it is validated the stem has the immediate prefix and suffix restored. (This should exclude all pronoun suffixes). This results in a word form (P).X.(S), where P is an optional prefix, and similarly for the suffix. This diminished surface word form is used to obtain, from the linguistic database for the verb type, the linguistic analysis for the word.

The rules are based on the root which is the 3rd person singular masculine in the past tense. The keys of the database contain the consonants of the root. Sixteen verb categories are defined in the database. Each verb category has a number of formalisms, defined as VC patterns, the surface word is allowed in the stem part for that particular verb category. Each VC formal pattern has in turn, a set of rules :

(prefix) + stem + (suffix) (2)

where the optional prefix and suffix are defined in the rules. One entry in the rule set for verbs of category type V11 is as follows

j ; u +VCCC+ verb(indicative, present, 2nd person plural) (3)

The first column has format
<prefix 1>, <prefix 2> ..<prefix i> ; <suffix 1>, <suffix 2>, ..

The meaning is any of the prefixes is valid. The ' ; ' is a delimiter when there are both prefixes and suffixes. The second column defines the surface stem for the current rule in the verb category and the valid affixes it can take. The stem definition is in terms of consonant and vowel positions. For a given root the stem consonants are identical to the root consonants in the sequence of the root consonants. The vowels in the stem do not necessarily have to be the same vowels as in the root.

For example consider the word under test *jiksruhomlok*. Initially all potential aggregate of consecutive consonants is examined as a potential verb stem. In particular for the consecutive consonants k, s, r, the database yields the verb *kiser*, in verb category V11.

Using (2) the affixes are *'ji'* and *'uhomlok'*. *'uhomlok'* is found in the initial analysis as being three valid suffixes - *'u', 'hom' 'lok'*.

'j' is found as a valid prefix.

The immediate affixes are returned so that the lexeme under test is now *'jiksru'*. Out of all the rules defined for verb category V11 to which *kiser* belongs, the rule definition (3) gives a valid outcome.

For Romance verbs a corresponding rule entry is as follows

n ; a,i +< >+ verb(indicative present first person singular) (4)

where the < > implies the root stem which remains the same in all verb forms for romance verbs.

For example given the word form *nippermettilek*, the consecutive consonants p,p,r,m,t,t find a database entry *ippermetti* under the verb category for romance verbs.

Using (2) *'n'* and *'ilek'* are found as valid affixes. The immediate affixes are returned and the rule (4) is then found to be satisfied.

RESULTS

The rules were tested with a series of sentences. Errors were considered to be of two types. These are no classification errors when either a function or a verb word was not classified, and misclassification when a verb word was classified as an adjective or noun, and viceversa. The results gave errors of the first type whenever the corresponding stem was not in the database. Errors of type 2 occurred in the case of one verb category, otherwise all function and verb words were classified correctly. The error occurs because in the verb category CVCV the derivational morphology between verb and noun or adjective involves only a change of vowels. For example *beda*, (he started); *bidu* (start). The analysis therefore yields both types as being possible valid answers for the word, though clearly a native speaker can distinguish between them. One of the future features to be added is to use a vowel tier on the present morphological analysis structure.

CONCLUSIONS

A rule based analysis suitable for a language with a mixed linguistic categories was designed. This was based on a separation of database and rule definitions enabling a quicker analysis as well as easier addition / deletion of rules based on a set of rule formalisms.

REFERENCES

1 J. Aquilina, "Maltese - English Dictionary Volumes 1 and 2", *Midsea Books Ltd.* 1987

2. P. Micallef, "A Text to Speech Synthesis System for Maltese", *unpublished Ph. D. thesis University of Surrey UK*, 1998