Crowd-sourcing evaluation of automatically acquired, morphologically related word groupings

Claudia Borg and Albert Gatt

University of Malta
claudia.borg / albert.gatt @um.edu.mt

Abstract

The automatic discovery and clustering of morphologically related words is an important problem with several practical applications. This paper describes the evaluation of word clusters carried out through crowd-sourcing techniques for the Maltese language. The hybrid (Semitic-Romance) nature of Maltese morphology, together with the fact that no large-scale lexical resources are available for Maltese, make this an interesting and challenging problem.

Keywords: computational morphology, segmentation, machine learning

1. Introduction

Automatic morphological clustering or grouping of words is an interesting task that can lead to the bootstrapping of a morphological analyser. This is useful for an under-resourced language such as Maltese, for which added difficulties arise due to its hybrid system that evolved from an Arabic stratum, a Romance (Sicilian/Italian) superstratum and an English adstratum (Brincat, 2011). The Semitic component follows a non-concatenative, root-and-pattern strategy familiar from languages such as Arabic and Hebrew, with consonantal roots combined with a vowel melody and patterns to derive forms, plus inflectional affixes for morphosyntactic features such as person, number and gender. By contrast, the Romance/Anglo-Saxon derivational component is concatenative (i.e. exclusively stem-and-affix based).

Table 1 displays examples of verb inflection and bound pronoun attachment for two verbs of Semitic origin (*niżel* 'to descend' and *żelaq* 'to slip') and one of Romance origin (*aċċetta* 'to accept'). The final row includes an example of the use of bound pronouns (glossed as ACC) to mark the direct object of a transitive verb, in addition to inflectional affixes.

	niżel √NŻL 'go down'	*żelaq* √ŻLQ 'slip'	*aċċetta* 'accept'
1SG	n-inżel	n-iżloq	n-aċċetta
2SG	t-inżel	t-iżloq	t-aċċetta
1PL	n-inżl-u	n-iżolq-u	n-aċċetta-w
2PL	t-inżl-u	t-iżolq-u	t-aċċetta-w
3SGM.3SGM-ACC	i-niżżl-u	i-żellq-u	j-aċċetta-h

Table 1: Root-based and stem-based morphology examples. All verbs are in the imperfective.

The work reported here is part of a broader effort to develop morphological resources for Maltese, which has received very little systematic computational treatment (Farrugia, 2008; Camilleri, 2013), despite a significant body of descriptive and theoretical research (Fabri, 2009; Schembri, 2006; Spagnol, 2011).

The aims of the present paper are twofold. It describes an unsupervised method for morphological clustering in Maltese, taking a wordlist extracted from the MLRS corpus of Maltese (Gatt and Čéplö, 2013)[1] as a starting point. The approach relies on a variety of relatedness heuristics which have been used successfully in previous work on other languages. In an effort to approach the problem in a completely unsupervised manner, we do not incorporate information about the historical origin (Semitic/root-based vs Romance/stem-based) of words. The second goal of this paper is to describe the evaluation method used for the output of our clustering method. While developments in this field have often been aided by the existence of gold-standard lexical resources (e.g. as used for the Morpho Challenge evaluations), no such resources currently exist for Maltese. Hence, we adopt a crowd-sourcing strategy, evaluating output with experts as well as non-experts, thereby also laying the grounds for the development of future gold standards. In section 2. we describe related work to this task, whilst in section 3. we provide a description of our technique. Section 4. reports on the evaluation carried out and the results achieved. Finally in section 5. we discuss the future directions of our work.

2. Related Work

Learning of morphological relations can be split into subtasks — segmenting words into morphemes, asso-

[1] http://mlrs.research.um.edu.mt

ciating labels to the individual morphemes, and grouping morphologically related words together (Roark and Sproat, 2007). The focus of this paper is on grouping together morphologically related words in Maltese through segmentation, orthographic and semantic similarity of words.

Goldsmith (2001) uses an unsupervised algorithm that separates stems from suffixes using minimal description length. Creutz and Lagus (2002; 2004; 2005) adopt a Maximum a Posteriori framework to segment words from unannotated texts; this has become the baseline evaluation for the Morpho Challenge competition (Kurimo et al., 2010). Schone and Jurafsky (2000; 2001) and Baroni et al. (2002) use orthographic and semantic similarity to detect morphologically related word pairs. Yarowsky and Wicentowski (2000) use a combination of alignment models with the aim of pairing inflected words. However this technique relies on part-of-speech, affix and stem information. Can and Manandhar (2012) create a hierarchical clustering of morphologically related words using both affixes and stems to combine words in the same clusters. In most of these approaches, evaluation relies on a gold standard, against which algorithms are compared using familiar metrics such as accuracy, precision, recall and f-measure.

For Semitic languages, the main issue in computational morphology tends to be that of disambiguation between multiple possible analyses. Habash and Rambow (2005) learn classifiers to identify different morphological features, used specifically to improve part-of-speech tagging. Snyder and Barzilay (2008) tackle morphological segmentation for multiple languages in the Semitic family and English by creating a model that maps frequently occurring morphemes in different languages into a single abstract morpheme.

Transitional probabilities are also used to determine word boundaries and discover affixes (Keshava and Pitler, 2006; Dasgupta and Ng, 2007), using an efficient Trie structure to calculate the probabilities. The technique is very intuitive, and posits that the most likely place for a segmentation to take place is at nodes in the trie with a large branching factor. The result is a ranked list of affixes which can then be used to segment words.

In the present paper, we give a brief overview of the clustering method which is based on some of the above techniques, before turning to the evaluation study, which is the main focus of the present paper.

3. Grouping Morphologically Related Words

Our starting point is a wordlist obtained from the MLRS corpus of Maltese, which contains 125 million tokens. Excluding function words, numbers, punctuation marks, proper nouns, determiners, foreign words and those words with a token count of less than 10, the resulting list has 67,434 word types. Function words were identified through frequency counts and checked manually, whilst the other words were identified through simple rule-based heuristics and further manual checking.

The clustering method described below is based on the identification of words with similar stems (whether these are integral stems of Romance origin, or are of Semitic origin and consist of root-and-pattern amalgams). In other words, our aim is to group together lexical items in the manner exemplified by the columns in Table 1. One of the challenges here, arising from the hybrid nature of the morphology, is that semitic stems frequently undergo allomorphy (Fabri, 2009); hence, even if an adequate segmentation process to strip inflectional affixes were present, it is not possible simply to group words based on stem identity. Thus, before comparing word types to determine their morphological relatedness, we first identify their likely affixes. We then apply clustering heuristics based on orthographic and semantic similarity on the resulting stems.

3.1. Segmentation-based Clustering

Words are segmented using a technique similar to Keshava and Pitler (2006), which determines where the most likely split should occur by calculating transitional probabilities of all possible boundaries in a word. The wordlist is modelled as a Trie structure which represents common sequences of characters in a single sequence, with branching whenever they diverge. Having $\alpha AB\beta$ be a word representation[2], the sequence $B\beta$ is considered as a likely suffix if (i) αA is itself a type in the wordlist; (ii) $P(A|\alpha) \approx 1$ (i.e. all occurrences of αA start with α, no substantial branching at this point); (iii) $P(\beta|\alpha A) < 1$, i.e. not all occurrences of αA end with β, substantial branching occurs. All potential splits are considered and scored according to their probability, resulting in a ranked list of prefixes and suffixes from which we use the top

[2]We use uppercase A, B for single characters, greek letters α, β for character sequences, and the possible boundary being examined between A and B. Hence, αA is the stem, and $B\beta$ the suffix. The system for finding prefix boundaries is exactly the same, with the words processed in reverse order.

10% (200 prefixes and 400 suffixes) to proceed with the segmentation process.

Words are segmented using the resulting list of affixes. However, the technique used can find more than one possible segmentation for a word. For instance, for *ipparkja* 'he parked', the method finds three potential segmentations: *i-pparkja*, *ip-park-ja* and *ippark-ja*. These complications may arise due to homographs in the wordlist. For instance, the stem *park* is both a part of the verb meaning 'to park' and the noun meaning 'public space'. Similarly, *spiċċa* 'to finish' might be wrongly segmented into *s-piċċ-a*, where the stem *piċċ* 'pitch' exists, but never takes the prefix *s-*.

A more subtle case is the presence of initial consonant gemination in Maltese as part of the inflection process: in this case, a perfective verb in the third person singular masculine duplicates the initial consonant (e.g. *park-ja* $\rightarrow$ *p-park-ja*) and, if it is preceded by a word ending in a consonant, takes the epenthetic vowel *i* (i.e. *i-ppark-ja*). As these examples illustrate, the problem is to determine when a word should *not* be segmented.

In order not to restrict the system, all possible segmentations are taken into consideration, and each potential stem is taken as the basis of forming word clusters which share the same stem; the resulting cluster is headed by that stem. This initial clustering process produces 21,381 clusters, with a word potentially present in multiple clusters.

3.2. Improvement of Clusters

Semantic relatedness and orthographic similarity of words is then introduced to improve the quality of the clusters and reduce their number. Latent Semantic Analysis (LSA) is a technique that analyses the relationship between words by comparing the context in which they appear. For example, the words *green* and *blue* would be expected to be surrounded by similar words since they are both colours. This is also plausible for morphologically related words, such as *blue* and *bluish*. LSA creates a vector representing the surrounding words for each word present in the corpus. It then computes the semantic similarity of two words by comparing the similarity in their individual vectors. The more likely two words are to appear in similar contexts, the more semantically related they are. In order to apply LSA, we chose to use the semantic space library implemented by Jurgens and Stevens (2010).

The purpose of applying semantic similarity between words is twofold. First, clusters with a high number of unrelated words should be disregarded completely, under the assumption that morphological relatedness should be correlated with semantic relatedness in all but the most unproductive morphological processes. In the latter case, of course, two words which may be morphologically related may have lost their semantic relatedness completely.

Second, clusters which might be morphologically related but were clustered separately, especially due to stem variation, should be considered for merging. For both cases we devised a metric that measures a cluster's *semantic cohesiveness*. The idea is that the more semantically related the words are within a cluster, the tighter or more 'compact' it is. Although semantic relatedness does not necessarily imply morphological relatedness, the clusters so far have been formed through the identification of potential stems, thus already limiting this measure to words which already have a strong orthographic similarity between them. Semantic cohesiveness is calculated by taking the standard deviation of the semantic similarity between the stem heading a cluster and every other word in the cluster. Let C_i be a cluster headed by stem s_i, and let $Sem_{s_i w_j}$ be the semantic relatedness of s_i and $w_j \in C_i$. The cohesiveness, σ_{C_i} is computed as

$$\sigma_{C_i} = \sqrt{\sum_{w_{1\ldots n} \in C_i} (Sem_{s_i w_j} - \mu)^2}$$

where μ is the mean pairwise semantic relatedness in C_i. The intuition here is that, the wider the dispersion within the cluster, the less cohesive it is.

So far, our initial clustering technique was based on the identification of a potential stem, and thus produces errors (e.g. by grouping together *iebes* 'hard' and *liebes* 'he is wearing') where morphologically unrelated words are clustered together due to high orthographic similarity. We posit that in the cases were words were wrongly clustered together, the semantic cohesiveness will be negatively impacted, whilst those clusters that contain only morphologically related words will have a tighter semantic cohesion value. We can then utilise this value to tackle the two problems identified above with the intention of retaining our clustering technique fully automated without any human intervention.

The merging of clusters is then carried out in two phases. First, we rank pairs of clusters that can be merged; this avoids the problem of merging clusters iteratively in a random order, which may result in C_i being merged with C_j, when a later cluster C_k would have been a better candidate to merge with C_i. Ranking depends on:

1. The semantic similarity between the two stems of the two candidate clusters, $Sem_{s_i s_j}$, where s_i is the stem heading C_i, and s_j is the stem heading

C_j. We use an empirically determined minimal threshold of 0.4.

2. The orthographic similarity between the two stems, a function of their minimum edit distance Med. We use a weighted version of Med, which favours character matches over deletions and insertions. This is important given the fact that words formed by non-concatenative processes with a common root will often share consonants, but have different patterns (e.g. *kiteb* 'he wrote' and *nkitbu* 'they were written', with common root $\sqrt{\text{KTB}}$). In our implementation, we assign costs as follows: match 4.0; insertion -6.0; deletion -6.0; substitution -12.0. The threshold is 0, meaning that the two stems have some matching characters.

3. The improvement in semantic cohesiveness should the two clusters be merged, where $Imp_{C_iC_j} = \sigma_{C_i \cup C_j} - \min(\sigma_{C_i}, \sigma_{C_j})$. If $\sigma_{C_i \cup C_j}$ is lower than the cohesion for either cluster, then there is an improvement (lower dispersion).

We combine all the three values described above into a single weighted measure, referred to as Combined Value ($ComVal$): $ComVal_{C_i \cup C_j} = \alpha Sem_{s_i s_j} + \beta Med_{s_i s_j} + \gamma Imp_{C_i C_j}$, where the weights were empirically determined as $\alpha = 0.2$, $\beta = 0.2$, and $\gamma = 0.6$.
The process of merging clusters involves three possible cases:

1. Two clusters are encountered that so far have not been merged. In this case, the merging is carried out.

2. Two clusters are encountered, one of which has already been merged, but the other hasn't. In this case, the system checks whether merging another cluster with the existing merged clusters would result in an acceptable improvement (using the same threshold described above). If yes, the merge is carried out.

3. Two clusters are encountered that have already been merged with other clusters. In this case the decision was not to consider these two for further merging, in order to avoid the creation of inordinately large word clusters created through successive merging.

The resulting operation leaves us with 4,524 clusters, a substantial reduction from the initial 21,381 clusters.

4. Evaluation and Results

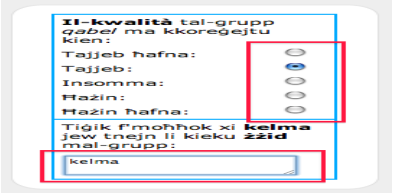
Since large-scale lexical resources are not available for Maltese, the evaluation must be carried out using human experts and non-experts. The evaluation focused on two groups of participants: (i) three linguists with postgraduate training (experts) and (ii) general native Maltese speakers (non-experts) sourced from the student population of the University of Malta, as well as through social networks. In total, 248 people visited the website, of which 151 carried out the evaluation.
The clusters presented for the expert evaluation were purposely chosen to represent a mixture of root-and-pattern and stem-and-affix based morphology. A total of 51 clusters were presented to each expert, 25 of which were shared amongst all experts so that the inter-annotator agreement between them could be measured. Experts were also given the opportunity to give textual feedback/comments on each of the clusters. The crowd sourcing evaluation had no limit on the number of participants. 300 clusters were randomly chosen, and our aim was to have at least 3 participants per cluster, especially for the calculation of inter-annotator agreement. A further 2 test clusters that were purposely put together to test any strategies used by participants in giving their judgements on clusters (for example, treating inflection and derivation differently).
The system to carry out the evaluation was built as a website[3] through which participants were provided with instructions in both written format and as a video, with an informal explanation of the concept of morphological relatedness through the use of examples. The evaluation itself presented a list of words together with the highlighted head word of a cluster, and participants could remove any words from this list (Figure 1(a)). They were also asked to rate the quality of the word cluster using a likert-type scale ranging from *tajjeb ħafna* 'very good' to *ħażin ħafna* 'very bad' (Figure 1(b)). Participants in the crowd-sourcing evaluation were exposed to clusters randomly; they could stop whenever they wished.

4.1. Removal of words

One of our main objectives in this evaluation is to have unrelated words removed from clusters so as to have clusters with only morphologically related words. From a quantitative perspective, analysing the percentage of words that were removed from a cluster provides insight into how well the clustering technique preforms. Table 2 divides the clusters into bins reflecting the range percentage of words removed, and dis-

[3] http://mlrs.research.um.edu.mt/cmexperiment/intro.php

(a) Removal of words



(b) Cluster rating (top) and addition of words (bottom)

Figure 1: Screenshot of the evaluation procedure

plays the percentage of clusters that fall into each bin. Separate columns are shown for Expert and Crowd-source evaluations and for the two Test clusters shared among all participants in the crowd-sourcing experiment.

Words removed	Expert	Crowd-source	Test
0%	54% (82)	56% (1025)	11% (26)
1 – 5%	1% (2)	1% (14)	0% (0)
5 – 10%	5% (8)	4% (79)	51% (122)
10 – 15%	3% (4)	3% (60)	7% (17)
15 – 20%	6% (9)	4% (77)	2% (37)
20 – 40%	15% (24)	16% (287)	12% (28)
40 – 60%	5% (7)	7% (137)	3% (8)
60 – 80%	9% (14)	1% (101)	0% (0)
over 80%	2% (3)	4% (68)	0% (0)
Total evaluations	153	1848	238

Table 2: Proportion of clusters with a given range of words removed by evaluators.

From the evaluations carried out, just over half (54% & 56% for experts and non-experts respectively) had no words removed. This is a positive indication given that the clusters were created fully automatically without any predetermined knowledge built into the program. At the other end of the scale, we have a very small percentage of evaluations where clusters had a rather large number of words removed. A manual analysis of these clusters indicates that there is general agreement between annotators that a particular cluster contained several unrelated words. For instance, one cluster that particularly stands out is for headed by the stem *ittra* 'letter'. This cluster contained several unrelated words due to *ittra* occurring as a substring in other unrelated words or having very close orthographic similarity — such as *tittraduċi* 'to translate', *ittratat* 'to be treated', *ittardja* 'to be delayed'. These types of errors were expected from this type of automatic system.

One of the problems is to achieve a balance between high precision (hence, stricter heuristics on establishing and merging clusters) and high recall (i.e. allowing a looser clustering involving more potentially related words). This balance tends to be achieved rather arbitrarily, based on empirically determined parameters. On the other hand, the overall percentages of words removed from the clusters are indicative that a rather acceptable balance was achieved with the majority of clusters having a reasonable percentage of words removed.

The test clusters provide us with further insight into how participants tackled the task. Very few evaluators appeared to distinguish between inflectional and derivational variants, for example, by removing derivations from a cluster and keeping only inflected forms of a stem. The vast majority preferring to leave both inflective and derived words in the cluster, a desirably outcome from our perspective. Each test cluster had at least one word which should have been removed from the given list to test for consistency. However in 11% of the evaluations participants did not remove any words. This can be due to not carrying out the task attentively, potentially reflecting a broader tendency on the part of these evaluators to make errors in removal. Thus it is important to look at the agreement between participants on which words were removed.

4.2. Inter-Annotator Agreement

We computed Inter-annotator agreement (IAA) between participants using Krippendorff's Alpha-Reliability (Krippendorff, 2011; Artstein and Poesio, 2008), a generalisation of several other reliability indices that take into account the likelihood of annotators exhibiting chance agreement and, crucially, is adapted to multiple raters (unlike, say, Cohen's Kappa). Mathematically, agreement is actually calculated by evaluating the observed disagreement (D_o) in relation to the expected disagreement (D_e) and subtracting this from 1, which would be full agreement. The basic formula is

$$\alpha = 1 - \frac{D_o}{D_e}$$

The coefficient ranges between 0 and 1, with 1 indicating full agreement.

We computed agreement over words, rather than over clusters, that is, for a given word that our algorithm places in a certain cluster, we are interested in the extent to which annotators agree that it should be in that cluster (based on the proportion of times an annotator removes it from the cluster). This makes agreement a binary classification: an annotator either keeps the word in the cluster, or not.

Let C be a cluster, consisting of words $w_1, \ldots, w_n$. For every evaluator, we represent the cluster as a vector v of binary values, so that $v_i = 1$ if word w_i was left in the cluster by the evaluator, and $v_i = 0$ if it was removed. The resulting matrix of vectors representing the evaluators' decisions for cluster C is then used to calculate the IAA using an implementation of Krippendorff's alpha[4] in R[5].

Description	Experts	Crowd-source
No. of Clusters	26	300
No. of Evaluations	78	1848
Avg. no. Evaluators	3	6.16
Avg. Agreement	0.908	0.598
Lowest Agreement	-0.0126	-0.166
Highest Agreement	1.0	1.0
Bins:		
Negative	4% (1)	23% (68)
less than 0.20	0% (0)	7% (21)
0.21 - 0.40	0% (0)	7% (20)
0.41 - 0.60	7% (2)	5% (16)
0.61 - 0.80	4% (1)	8% (24)
0.81 - 1.00	84% (22)	50% (151)

Table 3: Inter-Annotator Agreement

The first set of IAA results presented in Table 3 shows the number of clusters in the different evaluations carried out, the average number of evaluators per cluster and the average agreement achieved overall in the respective evaluation. The highest and lowest agreements are given, together with the percentage of clusters spread into bins according to the range of IAA achieved.

The expert group has the highest average agreement of 0.908, with 84% of the clusters having very high agreement between annotators. On the other hand, the average agreement of 0.598 for the crowd-sourcing evaluation is lower; this is expected, given gaps in linguistic knowledge and expertise of the general Maltese native speaker. The agreement between participants is however adequately high, especially when considering that 50% of the clusters have a very high agreement (between 0.81 and 1).

Somewhat surprisingly, there were some cases on which Krippendorff's alpha returned a negative value (meaning that evaluators are agreeing at a rate worse than chance). This is known to occur in case there is a minority of evaluators in the task who use one or more of the available categories, while the majority of evaluators do not use them. In fact, this result invariably arose when one evaluator excluded a small number of words from a cluster, while others didn't (i.e. there was only one evaluator whose decisions contained 0 values). In examining further the agreement of annotators in other clusters, we noted that there were quite a few cases where the negative or very low alpha was due to *outliers*, that is, cases where there is a strong agreement amongst the majority of the evaluators with one evaluator providing a different evaluation from the other participants. Thus we decided to recalculate IAA by excluding such outliers.

Outliers were detected as follows. We calculated the pairwise agreement between evaluators. An outlier was defined as an evaluator who consistently displayed low agreement with all other evaluators. If less than a third of the evaluators were outliers for a given cluster, then the IAA was re-calculated without the outliers. If more than a third of the evaluators were outliers, then all evaluators were included (as per the original IAA calculation), meaning that there is overall considerable disagreement between participants regarding which words should be removed. Overall, 210 (11%) evaluations were identified as outliers using this definition.

Table 4 shows the agreement results when such outliers are not included in the evaluation.

Description	Experts	Crowd-source
No. of Clusters	26	300
No. of Evaluations	77	1638
Avg. no. Evaluators	2.96	5.46
Avg. Agreement	0.948	0.897
Lowest Agreement	0.42	-0.156
Highest Agreement	1.0	1.0
Bins:		
Negative	0% (0)	2% (7)
0.00 - 0.20	0% (0)	2% (7)
0.21 - 0.40	0% (0)	2% (6)
0.41 - 0.60	7% (2)	2% (5)
0.61 - 0.80	4% (1)	9% (28)
0.81 - 1.00	89% (23)	82% (247)

Table 4: Inter-Annotator Agreement excluding Outliers

[4]The implementation used is part of the IRR library `http://rss.acs.unt.edu/Rdoc/library/irr/html/kripp.alpha.html`.

[5]`http://www.r-project.org/`

The most interesting change is in the crowd-source evaluation where the overall agreement has increased substantially, from an average of 0.598 to an average of 0.897. This is quite high considering that participants in this group in general have no linguistic knowledge regarding morphological relations.

4.3. Quality Ratings

Participants were also given the opportunity to rate a cluster in terms of its perceived quality. Although this is a rather subjective judgement, we were particularly interested in the correlation between the quality rating and the number of words removed for any given cluster. A *perfect* cluster is one which would have no words removed and given a high quality rating. A *bad* cluster would be one which would have a large number of words removed and given a low quality rating. Table 6 provides an overview of the quality ratings for each evaluation carried out, and the overall correlation coefficient between the quality rating and the percentage of words removed from a cluster. Correlations were computed using Spearman's rank-correlation coefficient, since one of our variables (rating) is ordinal.

Quality ratings	Experts	Crowd-source
Very Good	22% (34)	44% (810)
Good	35% (53)	30% (546)
Average	26% (40)	15% (272)
Bad	13% (20)	7% (128)
Very Bad	4% (6)	5 % (92)
Correlation:	0.737	0.788

Table 5: Quality ratings per evaluation, and Spearman's rank correlation coefficient between the quality rating and the percentage of words removed from a cluster. All correlations are significant at $p < .001$.

We note that overall, the majority of the clusters are rated 'good' or 'very good' (57% in the expert evaluation, and 74% in the crowd-source evaluation) in terms of their quality, and the overall correlation in both evaluations is rather high. It is interesting to note that the expert group were less inclined to give high quality ratings, whereas the non-expert group seemed more liberal with this rating. One possibility is that experts not only rate clusters based on the words they contain, but also implicitly consider the *completeness* of a cluster. This means that if a cluster contained few words that were all morphologically related, whilst a non-expert would probably give a high quality rating to such a group, an expert would give a lower quality rating in case the cluster had several missing, related words.

4.4. Hybrid morphology and the clustering technique

We now turn to one of the questions raised at the start of this paper, namely, the adequacy of the clustering technique for a hybrid morphological system in which different (concatenative and non-concatenative) processes are at work. Here, we briefly focus on the expert evaluation where, as mentioned above, the clusters were selected to be roughly balanced between the two processes.

	#Clusters	Average quality	#Intact
Concatenative	80	2.3	49 (61%)
Non-concatenative	73	2.55	33 (45%)

Table 6: Overview of performance of the clustering technique on different morphological processes.

Dividing the clusters according to their morphological processes, we have 80 concatenative clusters, and 73 non-concatenative clusters. Analysing the results, the main difference between the two groups is the percentage of clusters that had no words removed. 61% of the concatenative clusters had no words removed, against 45% of the non-concatenative clusters. This indicates that the technique performs better on the concatenative processes, nonetheless providing adequate results for non-concatenative processes. The percentage of clusters that had more than a third of their words removed is approximately the same (19% and 18% respectively), which shows that the system fails in both cases at approximately the same rate.

5. Conclusion

This paper briefly described a fully automatic technique that groups together morphologically related words using a combination of probabilities, semantic and orthographic similarity of words. The clusters were evaluated by linguists and native speaking non-experts. The results are rather promising, providing a basis for the bootstrapping of a morphological analyser. The evaluation dataset will be made available since it can be a useful resource for further analysis.

The clusters, together with their underlying segmentation, can also be used as an input to a semi-supervised machine learning technique in an attempt to introduce labels to the segments, as part of the process of developing a fully fledged morphological analyser. Given the existence of manually evaluated clusters, it is also conceivable that they might provide a basis for supervised clustering methods.

The development of a morphological analyser for Maltese remains a challenge due to its hybridity and

the lack of resources against which to evaluate, though The latter scenario is slowly improving. We plan to integrate new resources into semi-supervised techniques to incrementally learn labels for morphological segments. Finally, we also plan to evaluate our techniques on other languages.

6. Acknowledgements

The research work disclosed in this publication is partially funded by the Malta Government Scholarship Scheme grant.

7. References

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555—596.

Baroni, M., Matiasek, J., and Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*, MPL '02, pages 48–57. Association for Computational Linguistics.

Brincat, J. M. (2011). *Maltese and other Languages*. Midsea Books, Malta.

Camilleri, J. J. (2013). A Computational Grammar and Lexicon for Maltese. Master's thesis, University of Gothenburg, September.

Can, B. and Manandhar, S. (2012). Probabilistic hierarchical clustering of morphological paradigms. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 654–663, Avignon, France, April. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*, MPL '02, pages 21–30. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2004). Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 43–51. Association for Computational Linguistics.

Creutz, M. and Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 106–113.

Dasgupta, S. and Ng, V. (2007). High-performance, language-independent morphological segmentation. In *NAACL HLT 2007: Proceedings of the Main Conference*, pages 155–163.

Fabri, R. (2009). Stem allomorphy in the Maltese verb. In *Ilsienna - Our Language*, volume 1, pages 1–20, Germany. Brockmeyer Verlag.

Farrugia, A. (2008). A computational analysis of the Maltese broken plural. Bachelor's Thesis, University of Malta.

Gatt, A. and Čéplö, S. (2013). Digital Corpora and Other Electronic Resources for Maltese. In *Proceedings of the Corpus Linguistic Conference*, Lancaster, UK.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27:153–198, June.

Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 573–580, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jurgens, D. and Stevens, K. (2010). The S-Space Package: An Open Source Package for Word Space Models. In *System Papers of the Association of Computational Linguistics*.

Keshava, S. and Pitler, E. (2006). A simpler, intuitive approach to morpheme induction. In *PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, pages 31–35.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability.

Kurimo, M., Virpioja, S., Turunen, V., and Lagus, K. (2010). Morpho challenge competition 2005–2010: evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, SIGMORPHON '10, pages 87–95. Association for Computational Linguistics.

Roark, B. and Sproat, R. (2007). *Computational Approach to Morphology and Syntax*. Oxford University Press.

Schembri, T. (2006). The Broken Plural in Maltese: An Analysis. Bachelor's Thesis, University of Malta.

Schone, P. and Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, pages 67–72. Association for Computational Linguistics.

Schone, P. and Jurafsky, D. (2001). Knowledge-free induction of inflectional morphologies. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–9. Association for Computational Linguistics.

Snyder, B. and Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, Ohio, June. Association for Computational Linguistics.

Spagnol, M. (2011). *A tale of two morphologies. Verb structure and argument alternations in Maltese.* Ph.D. thesis, University of Konstanz.

Yarowsky, D. and Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 207–216, Stroudsburg, PA, USA. Association for Computational Linguistics.