Cross-lingual part-of-speech tagging for Maltese

Keith Lia & Héctor Martínez Alonso
University of Copenhagen
dhl107@alumni.ku.dk

ABSTRACT

We present a cross-lingual part-of-speech tagger which makes use of word embeddings to predict the part-of-speech tags for Maltese words after training on labeled English data. We first build a dictionary of bilingual word pairs, which we then use to create a unified bilingual corpus consisting of sentences with only Maltese words, only English words and a mixture of both. From this corpus we then induce cross-lingual word embeddings to use as augmenting features in the tagger. We conduct several experiments with different system parameters and compare them to our baseline systems that do not incorporate the word embeddings. We show that state-of-the-art approaches can be generalized to other languages and show how the results can be improved when word-order difference is taken into consideration.

1. INTRODUCTION

Maltese is an under-resourced language (Rosner et al., 2012). With about 520k native speakers, it is the national language of the Maltese archipelago. While it is a Semitic language–indeed, the only one in the European Union—, its lexicon is heavily influenced by Italian and English. Moreover, English enjoys co-officiality in Malta, and a lot of the written language use of the archipelago is still in English, in spite of the preponderance of Maltese as spoken language.

Better language technology can improve the presence of under-resourced languages in digital forms. While there exist initiatives such as the Maltese Language Resource Server[1] and the Maltese Language Software Services,[2] the support for many language-processing tasks is fragmentary.

This low-resource scenario justifies boostrapping resources and reapplying technology, previously developed for other languages, for Maltese. We propose applying cross-lingual methods to re-use training data for a source language (typically English) to predict linguistic properties in Maltese text.

[1] http://mlrs.research.um.edu.mt
[2] http://metanet4u.research.um.edu.mt

Cross-lingual part-of-speech tagging for Maltese Keith Lia and Héctor Martínez Alonso
.

In this article, we focus on the fundamental task of part-of-speech (POS) tagging.

While Arabic is the major language that is most related to Maltese and would be thus the more natural choice for cross-languistic processing, the difficulties in obtaining well-aligned translation pairs for Arabic and Maltese force us to use English as a source.

This article's main contribution is an adaptation to Maltese of a current state-of-the-art cross-lingual approach for POS tagging. Moreover, we propose a strategy to compensate for the difference in word order, which is a major cause of error in cross-lingual transfer of part of speech taggers. Section 3 describes how we extend Gouws and Søgaard (2015) to compensate for word-order difference when training with English as source language.

A common POS tagset is necessary for this task. We provide a conversion from the full Maltese tagset (Gatt and Céplö, 2013) to the 17-tag POS tagset used by the Universal Dependencies treebanks, cf. (Nivre, 2014). We also make available the embeddings used for the experiments at hand.[3]. This work is, to the best of our knowledge, the first cross-lingual experiment with Maltese as a target language.

2. RELATED WORK

Cross-lingual language processing techniques allow using data from a source language to predict linguistic traits (parts of speech, dependency structures) of another target language (Yarowsky et al., 2001; Søgaard et al., 2015; Fossum and Abney, 2005). Low-resource languages are ideal targets for cross-lingual techniques (Agić et al., 2015). These approaches rely very often on a shared bridge representation between source and target language.

Using bilingual dictionaries as bridging method has been addressed among others by Das and Petrov (2011), who use word-aligned text to automatically create type-level tag dictionaries. Täckström et al. (2013) expand the approach with type- and token-based distant supervision, building on the cross-lingual clustering approach of Täckström et al. (2012). Xiao and Guo (2014) perform cross-lingual dependency parsing using neural networks to build bridge representations (i.e. *word embeddings*) that pivot on bilingual lists of translated words. In a similar vein, Gouws and Søgaard (2015) use non-parallel corpora of source and target languge, plus a translation list they use to randomly generate a third mixed-language corpus. The three corpora are used together to fit word embeddings for both languages.

[3] https://github.com/KLia/MalteseXLingPosTag

They use their method on the language pair, both for POS and supersense tagging. Our approach is an extension of their method to compensate for word-order differences.

3. METHOD

This section describes the data used for the cross-lingual experiments, and how we extend the bilingual-word pair translation approach.

3.1 Data

Table 1 shows the number of sentences, word tokens and word types respectively of the corpora used in this work. For English, we use unlabeled data from the American National Corpus (Ide and Macleod, 2001) and from the WaCky Corpus (Baroni et al., 2009), and labeled data from Universal Dependencies.[4] For Maltese, we use labeled and unlabeled data from the Maltese Language Resource Server and Gatt and Céplö (2013). We isolate a development section for each language. The Maltese development section is used for feature extraction and error analysis (cf. Section 4.1.1 and 4.3).

		sentences	tokens	types
Unlabeled	EN	4M	97M	1M
	MT	4M	124M	891k
Labeled	EN	510k	13M	171k
	MT	265k	9M	159k
Development	EN	16k	254k	23k
	MT	15k	387k	31k

Table 1: Corpus size statistics.

3.2 Mixed-language corpus

In order to create a bridge between the two languages, we followed the methods described in (Xiao and Guo, 2014; Gouws and Søgaard, 2015) to make use of bilingual dictionaries to obtain a list of translated word pairs. We use this list to create a mixed-language corpus C' by randomly replacing words from the source and target language corpora, pooled together as C.

In the first attempt, we followed Gouws and Søgaard (2015) and used Wiktionary to obtain word pairs, but the coverage (1.10% types and 28.03% tokens of the Maltese corpus) was insufficient. Instead, we opted for building the word pair list using the Google Translate API.[5] We discarded using Arabic as a source language, because this translation system uses English as a pivot language when translating between Arabic and Maltese, thus making translated word pairs less reliable.

We translated the top N most-frequent items from the source unlabeled corpus, i.e. unigrams or bigrams, depending on the system (cf. Section 4) and obtained the corresponding word pair lists. We set $N = 100,000$.

For a certain word-pair list, we apply it to a corpus C that is made up of the concatenation of the English and Maltese unlabeled data, underlined in Figure 1. For the Maltese section of C, we randomly replace an item (unigram or bigram) by its English translation with a 20% probability, thus

[4] http://universaldependencies.github.io

[5] http://pythonhosted.org/goslate/

$C' =$

$\underline{MT}$
$\underline{EN}$
$MT \triangleright EN$
$EN \triangleright MT$

Figure 1: Language composition of the mixed-language C' corpora.

yielding a text that is about 80% Maltese. The process operates left to right, sentence-wise. If a word has been chosen for translation but is not covered by the translation pair list, the system chooses the next one, until a translation is found. We repeat this operation for English, replacing 20% of the items with their Maltese translation. The ideal proportion of original and translated text in the mixed-language section of C' is a potential parameter, but we used a fixed 80-20% rate because we expect a more even proportion to be much noisier and yield worse embeddings.

We applied the same item-translation procedure to a) unigram, b) bigrams, and c) unigrams and bigrams simultaneously. Thus, we create three variants of C', namely C'_u, C'_b and C'_{u+b}. Figure 1 shows a graphical representation of C', made of the Maltese corpus, English corpus and the mixture of both. Notice that, while C has 8M tokens (4M from Maltese and 4M from English), any of the three variants of C' has about twice that amount.

3.3 Local word order difference

Maltese and English have different syntax. For example, adjectives in English precede nouns whereas they follow nouns in Maltese. Therefore, translating the words '*black + cat*' individually would give an ungrammatical result, while translating the full bigram would produce the correct form.

Translating bigrams as a unit is a way of circumventing the problem of local (as opposed to long-distance) word order difference. Moreover, not all unigram translations are one-to-one. For instance, English subject is mandatory, while Maltese subject is not (*I enter = nidħol*). Symmetrically, Maltese possessive determiner are clitics attached to nouns instead of independent tokens (*my mom = ommi*), just as pronoun clitic attach to particles (*on me = fuqi*). We counted how many English uni- and bigrams in our dictionary were translated to Maltese unigrams, bigrams and n-grams of higher sizes. We report the figures in Table 2. The numbers are given as percentages.

The number of English bigrams translated into Maltese unigrams tallied up to 20.78%, around one-fifth all of translated phrases. On the other hand, Maltese unigrams were translated into bigrams more often than English unigrams; when translating from a mostly isolating language like English to a more inflected language like Maltese, we would benefit from considering bigrams in addition to unigrams as this would generate better-quality mixed-language corpora.

Word order variation when transferring across languages is not explicitly tackled in the work of Gouws and Søgaard (2015). In order to estimate the relevance of this variation, we compare the distribution of POS bigrams across the languages tags as shown in Table 3. We compare our source and target languages with Danish (a target language for Gouws and Søgaard) and Arabic, the typologically clos-

		Size of translated n-gram			
		1	2	3	4+
Unigrams	**EN▷ MT**	85.54	13.36	0.92	0.18
	MT▷ EN	56.15	33.80	6.13	3.92
Bigrams	**EN▷ MT**	20.78	72.71	5.72	0.79
	MT▷ EN	12.62	81.07	5.76	0.51

Table 2: Translation-size proportions in %.

est major language for Maltese, which could make a better source if translated word lists were obtainable in a reliable manner. We expect more internal consistency between Germanic and between Semitic languages than in cross-family comparisons. By comparing bigrams between two target languages of different families, we highlight the risks of inducing spurious POS sequences when applying cross-lingual models without taking word order in consideration.

Danish	English	Maltese	Arabic
Noun Adp	**Noun Adp**	**Adp Noun**	**Adp Noun**
Adj Noun	**Adj Noun**	**Noun Adj**	**Noun Adj**
Det Noun	Det Noun	Det Noun	Noun Noun
Det Adj	Adp Det	Verb Adp	Noun Adp
Adp Noun	**Pron Verb**	Aux Verb	Adj Adp

Table 3: Top 5 POS bigrams per language.

We can see that there are some immediate differences in tag order between the three corpora. In particular that adjectives precede nouns in English and Danish, but not in Maltese.In English and Danish, the bigram Pron Verb appears in the fifth and thirteenth position of all bigrams respectively, while in the Maltese corpus it appears at the 40th position . Similarly, Pron Aux is much more common in English and Danish, occurring at positions 14 and 22 respectively. In Maltese, this pair of POS tags occurs at position 70 , because Maltese is a pro-drop language, whereas Germanic languages have obligatory subject.

Finally, we present the Spearman correlation coefficient for the bigram ranks between languages in Table 4. Alongside the Overall rank, we split the POS bigrams into *Content* and *Function* to see where the biggest difference lies. For *Content*, all the bigrams containing the tags Det, Adp, Part, Conj, Sconj, Num, Punct and Sym were removed, while for *Function*, only bigrams containing at least one of these tags were considered.

As expected, English and Danish are more correlated than Maltese and English, with *Content* bigrams having the biggest difference in correlation rank. The table also shows that for Maltese, bigram tags for content words are less correlated than bigrams that include function words.

We use the variation in number of tokens when translating (Table 2), and the language-family similarity in the POS bigram distribution (Tables 3 and 4) as supporting evidence for the use of a bigram dictionary in order to narrow the gap between the two languages. Furthermore, a bigram-translation approach potentially improves over the tendency

	Overall	Content	Function
DA-EN	0.85	**0.84**	0.84
MT-EN	0.79	0.78	0.83
MT-AR	0.80	**0.83**	0.72
EN-AR	0.77	0.78	0.73

Table 4: Bigram rank correlation across languages.

of Wiktionary to have better coverage for lemmas than for the rest of the forms, which is a challenge when e.g. dealing with Maltese inflectional morphology. In addition to noun and verbal morphology, the form of the definite article is determined by the first phoneme of its introduced noun, e.g. '**il**-lapes' ("the pencil") vs. '**iż**-żunżana' ("the bee"). Cf. Rosner et al. (2012) for a summary on Maltese morphology.

4. EXPERIMENTS

For our experiments, we used a structured perceptron[6] to perform POS-tagging on English as the source language and Maltese as the target language. Table 6 shows all the different systems used for our experiments. These include two baselines and 12 systems which use our cross-lingual word embeddings, plus two extra systems with type constrains. The simple delexicaled baseline (BL) uses only the shape and count features, while the extended baseline (BL-E) also uses the word translation feature. The rest of the systems used the shape and count and the embedding features.

Table 6 lists, for each system that uses WE, its variant of C' , be it unigram (u), bigram (b) or both (u+b) (cf. Section 3.2). The Extra column shows whether a system uses simple word translation (wt) or type constrains on the hyphenated prefixes (pc) or on all stop words (sc), cf. Section 4.3.

4.1 Features

For any word w, we extract the following token features.

4.1.1 *Shape and count*

These features describe simple properties of w and are commonly used in a cross-lingual setup. Some features describe certain POS, like *punct* for the Punct tag, *num* for Num and *cap* for Propn. The feature *stop* covers frequent words in a language, commonly Adp, Conj, Det, Part and Sconj. We used the English stoplist from NLTK.[7] A Maltese version does not exist, and we extracted all the words from our development set (cf. Section 3) with these five tags.

4.1.2 *Word translation*

The **trans** feature in Table 5 shows the word translation feature of the extended baseline. We used a feature to make an alignment between the words in different languages by using a bilingual word translation dictionary (cf. Section 3.2) to translate w from the original language into our target language. The bilingual dictionary is only used to fill out the value for the training set containing the English corpus, whereas the word token itself is used in the test set containing the Maltese corpus.

4.1.3 *Word embeddings*

[6] https://github.com/coastalcph/rungsted
[7] http://www.nltk.org/

Feature	Description	Value
cap	Whether w is capitalized	True/False
pos	The relative position of w in the sentence, according to 3 equal-sized windows	{1,2,3}
len	Length of w in characters	$\mathbb{N}$
punct	Whether w includes punctuation	True/False
alnum	Whether w word contains alpha-numeric symbols	True/False
alpha	Whether w contains only letters	True/False
digit	Whether w contains only digits	True/False
stop	Whether w contains is a stop word	True/False
trans	**Train set:** The Maltese translation of w **Test set:** w itself	string
emb_n	The n-th position in the embedding vector for w	$\mathbb{R}$

Table 5: Shape and count features.

We made use of word embeddings as features in our tagger. Table 5 shows an abstracted view of each word-embedding feature. We calculate embeddings on all three variants of C' described in Section 3.2. We preprocess the corpora by lowercasing the text, removing punctuation and one-token sentences. We only calculate embedding vectors for words that appear at least five times. The word embeddings were induced using hierarchical soft-max for 100 dimensions on word2vec.[8] Each component of the embedding vector is used as a feature, creating 100 new features—one for each dimension. We experiment with two sampling methods, namely continuous bag-of-words (CBOW) and skip-gram, and with two different window sizes (2, 5). Cf. Mikolov et al. (2013) for a discussion on sampling methods.

4.2 Results

We provide the F1 for each system in Table 6. The first baseline is a delexicalized tagger using only the cross-lingual features. The second extended baseline is a re-lexicalized baseline with an additional word translation feature and gives a better performance than our original baseline. All the embeddings systems beat the extended baseline significantly at $p < 0.05$ on a bootstrap test.

The system with the highest micro-averaged F1 is C5B, namely CBOW on C'_b with a window size of 5. We also compare the results with type-contrained distant supervision on the best system (C5B), which we describe in Section 4.3.

4.3 Performance across POS tags

In this section we describe the results for the systems that use the word embedding features. We only report the results for the systems using the CBOW models as these outperformed the Skip-gram systems. Figure 2 shows a system-wise comparison of the F1 for each POS tag, except SYM and

[8] https://radimrehurek.com/gensim

	Extra	WE unit	WE sample	WE window	F1
baseline					0.32
baseline-Ext	wt				0.53
cbow-2-uni		u	cbow	2	0.62
cbow-5-uni		u	cbow	5	0.64
cbow-2-bi		b	cbow	2	0.61
cbow-5-bi		b	cbow	5	**0.65**
cbow-2-unibi		u+b	cbow	2	0.62
cbow-5-unibi		u+b	cbow	5	0.62
sg-2-uni		u	sg	2	0.61
sg-5-uni		u	sg	5	0.54
sg-2-bi		b	sg	2	0.64
sg-5-bi		b	sg	5	0.59
sg-2-unibi		u+b	sg	2	0.55
sg-5-unibi		u+b	sg	5	0.56
C5B+P	pc	u+b	cbow	2	**0.75**
C5B+S	sc	u+b	cbow	2	**0.81**

Table 6: The 12 systems used in our experiments, with the features used in each, plus two extra systems with type constrains.

X. Each box's color represents the system with the highest score for a given POS.

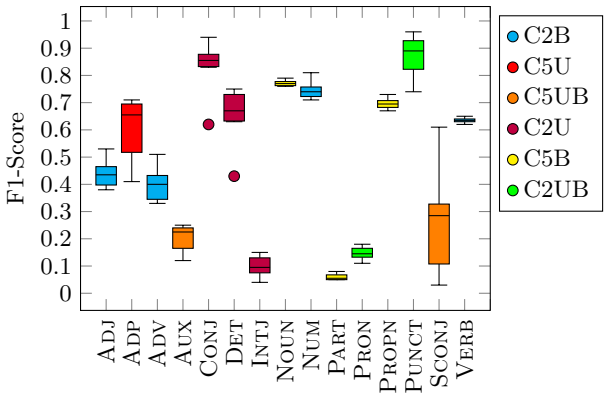

Figure 2: F1-score using CBOW word-embeddings. The colours represent the system with the highest score for that particular POS tag.

As per the content POS, nouns and verbs had very little variance across systems, although systems using bigrams achieve higher results. On average, both classes of tags boosted their F1-score by 0.10 over the extended baseline. Bigram systems perform better for adjectives because of the improved word-order information of the translation.

For function POS, the tag SCONJ showed a great variation between the n-gram systems, with C2U performing as low as 0.03. The C5UB system achieved the best score for SCONJ at of 0.61 as a consequence of the more strict contexts generated by translating bigrams and unigrams. CONJ and DET also show variation but perform much better across systems.

Determiner performance shows remarkable variation among the systems, with unigram systems getting a better score for this class. This behavior contradicts our expectations about the phonetic assimilation of determiners to nouns (cf. Section 3.3), which should be best captured by bigram transla-

tions. In general, DET is misclassified as the PART or ADP, with the bigram systems showing more errors. We compared word embedding vectors from C5B and *C5U* for determiners with vectors for adpositions and particles as shown in Table 7 using cosine similarity.

System	DET	Word	POS	Similarity
5bi	il-	id-	DET	**0.24**
5bi	il-	tal-	PART	0.55
5bi	il-	bil-	ADP	0.48
5uni	il-	id-	DET	**0.41**
5uni	il-	tal-	PART	0.55
5uni	il-	bil-	ADP	0.44

Table 7: Comparison of DET with ADP/PART.

The two determiner vectors from the unigram system have a higher similarity score than the same determiners from the bigram system. In both systems, the vectors for *il-* is more similar to the vectors for *tal-* and *bil-*, which belong to different classes, however in the unigram system, the difference between similarity score is much less pronounced. The distributional information encoded in the vectors does not differentiate between DET, ADP and PART, but merely captures that these words are prenominal.

Given the performance variation in the function POS, we apply two variants of type constrains, one for hyphenated prefixes (C5B+P), and one for all known stop words (C5B+S). For the former, we extracted the hyphenated words tagged as ADP, DET and PART from the development set, for a total of 72 word types. In this subset of our stoplist, each word type had strictly one possible POS without any ambiguities, so we develop a post-processing stage where every token in this list would be assigned the correct POS tag. These two systems occupy the lower rows of Table 6.

We present the F1-scores for all tags achieved by our best overall system C5B in Table 8, compared to its prefix type-constrained variant (C5B+P), its unigram counterpart (C5U), and the extended baseline. C5B beats BL-E for all tags except ADV and PRON, where the high recall does not make up for low precision, e.g. 62.25% of all tokens tagged as PRON where in fact SCONJ. Even though the micro-averaged F1 of C5B is very similar to C5U, the unigram system is better at identifying function words, and we prioritize the disambiguation of content POS, because it is easy to boost function-POS performance using type constrains.

Tag	BL-E	C5U	**C5B**	C5B+P
ADJ	0.34	0.42	**0.47**	0.51
ADP	0.51	**0.71**	0.68	0.74
ADV	0.59	0.41	**0.44**	0.50
AUX	0.14	0.24	0.24	0.35
CONJ	0.83	**0.87**	0.83	0.86
DET	0.48	**0.74**	0.70	0.85
NOUN	0.65	0.77	**0.79**	0.82
NUM	0.25	0.72	**0.75**	0.77
PART	0.03	**0.20**	0.04	0.30
PRON	0.26	0.15	**0.17**	0.24
PROPN	0.66	0.67	**0.73**	0.77
SCONJ	0.05	**0.33**	0.28	0.44
VERB	0.50	0.56	**0.64**	0.70

Table 8: POS-wise F1 scores between systems.

4.3.1 *Performance across domains*

Our Maltese test corpus is made up of several domains: academic, law, literature, newswire, religion, speeches, web (wiki), web (general) and miscellaneous. We evaluate the text from each Maltese domain individually.

In Table 9, we list the results for each domain evaluated both on the extended baseline and on our best system *C5B*. The OOV_{WE} (word-embedding out-of-vocabulary rate) column indicates the percentage of word types in each domain that did not receive an embedding vector—and thus embedding features—because their frequency was under 5.

Domain	Tokens	BL-E	C5B	OOV_{WE}
Academic	176k	0.47	0.60	2.6
Law	1M	0.54	0.66	0.7
Literature	267k	0.56	0.67	0.7
Newswire	6M	0.56	**0.68**	0.8
Religion	459k	0.45	0.59	4.3
Speeches	20k	0.54	0.63	1.2
Web-wiki	559k	0.52	0.63	4.7
Web-general	656k	0.47	0.59	5.3
Miscellaneous	139k	0.52	0.63	0.9

Table 9: F1-Scores for each of the Maltese domain evaluated on the baseline and *C5B*

Newswire is the best-performing domain. Besides being the most represented in our unlabeled data and development data, it is present in the domains of the training data, and, it as it has the highest word-pair dictionary coverage (Sec cf. 3.2), which also gives it a low OOV_{WE}.

Embedding coverage correlates with performance. Indeed, the domain with highest OOV_{WE} (Web-general) has the lowest F1, which is shared by Religion , also with a high OOV_{WE}. However, domains with better coverage match other lower-performing domains; Academic and Miscellanous are not far from Web-general, in spite of a low OOV_{WE}. We attribute this result to the variation in POS sequences across domains, and not only to the words themselves.

5. CONCLUSIONS

We have shown a method to apply the cross-language method of Gouws and Søgaard (2015) with a bigram-translation approach to compensate for local word-order differences. Our best system with type constrains achieves an F1 of 0.81, substantially over the baseline.

For any variant of the mixed-language corpus C', we determine that CBOW outperforms skip-gram sampling. This behavior is understandable given that skip-gram embeddings are more topical and contain less word-sequence information, which is a key factor for POS tagging.

Our type-constrain method is essentially a post-processing technique. A decoding-step distant-supervision method like the one used by Täckström et al. (2013) would further improve our results without requiring more language resources.

Our method can be applied to other cross-lingual tasks where the the target shows sensitive word-order differences with regards to the source language. Extending this approach to dependency parsing could be used to bootstrap the development of a Maltese treebank.

Moreover, as regards Maltese, improving POS tagging also allows raising the bar of subsequent NLP applications that benefit from accurate POS prediction.

Acknowledgements

This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Agić, Ž., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

Fossum, V. and Abney, S. (2005). Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *Natural Language Processing–IJCNLP 2005*, pages 862–873. Springer.

Gatt, A. and Céplö, S. (2013). Digital corpora and other electronic resources for maltese. In *Proceedings of Corpus Linguistics Conference, University of Lancaster*.

Gouws, S. and Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL.

Ide, N. and Macleod, C. (2001). The American national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, volume 3.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nivre, J. (2014). Universal dependencies for Swedish.

Rosner, M., Joachimsen, J., Rehm, G., and Uszkoreit, H. (2012). *The Maltese language in the digital age*. Springer.

Søgaard, A., Agić, Ž., Martínez Alonso, H., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722, Beijing, China. Association for Computational Linguistics.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487. Association for Computational Linguistics.

Xiao, M. and Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. *CoNLL-2014*, page 119.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.