Converting Afrikaans to Dutch for technology recycling

Suléne Pilon
School for Languages
North-West University (VTC)
Vanderbijlpark
Sulene.Pilon@nwu.ac.za

Gerhard van Huyssteen
Centre for Text Technology (CTexT)
North-West University (PC)
Potchefstroom
Gerhard.VanHuyssteen@nwu.ac.za

Liesbeth Augustinus
Centre for Computational Linguistics
Katholieke Universiteit
Leuven
Liesbeth.Augustinus@gmail.com

Abstract—**HLT resource development for a resource scarce language (*L2*) can be expedited by recycling existing technologies for a closely related language (*L1*). To improve the success of *L1* technologies on *L2* data, one can convert *L2* data to make it appear more *L1*-like. We explore this possibility by developing an Afrikaans-to-Dutch lexical conversion module and using it as pre-processing step before applying a Dutch part of speech tagger to Afrikaans data. The accuracy of the Dutch tagger increased from 62.6%, when tagging raw Afrikaans data, to 80.6% when tagging converted Afrikaans data. We therefore conclude that, at least in the case of Dutch and Afrikaans, the use of lexical conversion as a pre-processing step for technology recycling merits further investigation.**

Keywords-technology recycling; lexical conversion; Part of speech tagging; Afrikaans; Dutch

I. INTRODUCTION

To adapt/re-engineer existing technologies for language *L1* to a closely related, resource-scarce language *L2* is a strategy that could be adopted to fast-track the development of resources for *L2*. The rationale behind this process of technology transfer is that if the languages are similar enough, it should be faster and cheaper to adapt *L1* technologies to *L2* than to develop *L2* technologies from scratch [1]. In [2] we have argued that Dutch and Afrikaans are similar enough, and that it would thus be easier and quicker to adapt existing Dutch technologies to Afrikaans, rather than to develop Afrikaans resources from scratch.

To improve the efficiency of *L1* technologies on *L2* data, and thereby further reduce the need for manual intervention, an extra layer of processing could be added during which *L2* data is manipulated to appear more *L1*-like. The nature of this processing would depend on the differences between the languages in question, and could include syntactic re-ordering, the splitting of morphemes, or some form of machine translation. Given the differences between Afrikaans and Dutch (see [2]), for example, a lexical convertor could be used to convert Afrikaans lexemes to Dutch before a Dutch POS tagger is applied to the data. Even though this conversion would not yield a good Dutch translation, the fact that the data appears more Dutch-like raises the question whether this will have a positive effect on the success of the tagger. This is the research question we would like to address in this paper, and we describe the development and application of such an Afrikaans-to-Dutch convertor (A2DC).

We first give an overview of previous work in Section 2, after which A2DC is explained in Section 3 and a word-level and sentence-level evaluation of A2DC is described in Section 4.1. Since A2DC was developed to expedite and improve technology recycling between Afrikaans and Dutch, we then use a Dutch POS tagger to annotate unconverted and converted Afrikaans data to evaluate the contribution that the conversion step makes in the recycling process. This experiment is described in Section 4.2. Section 5 concludes and gives a view to future work.

II. PREVIOUS WORK[1]

To facilitate the development of lexical convertors, we developed a lexical conversion module, which consists of a language independent algorithm and four language specific resources, *viz.* a specialised bilingual dictionary (Lex.LangIn-LangOut.txt), a target language lexicon (Lex.LangOut.txt), language-pair specific morpheme conversion rules (MorphRules.txt), and language-pair specific grapheme-to-grapheme conversion rules (G2GRules.txt). Since Dutch is generally considered to be morphologically more complex than Afrikaans, we first used this conversion module to develop a Dutch to Afrikaans lexical convertor (D2AC) to serve as proof-of-concept for rule-based lexical conversion between Afrikaans and Dutch [2].

In [2] we reported an accuracy of 71% on word-level and a BLEU score of 0.2519 in a small-scale evaluation on running text. After some minor improvements to D2AC, which includes expansion of the bilingual lexicon and rule optimisation and re-ordering, we obtained an accuracy of 71.8% in the word-level evaluation (see [3]). To illustrate the differences between $D2AC_{2009}$ [2] and $D2AC_{2010}$ [3], the sizes of the language dependant components of each of the convertors are given in Table 1.

We also conducted a sentence-level evaluation to compare the convertor to the Dutch-Afrikaans Google Translate (GT). The results of this evaluation are shown in Table 2. It is not surprising that GT achieves a higher BLEU score than $D2AC_{2010}$ in the automatic evaluation, since $D2AC_{2010}$ is only

[1] Please see [2] for a detailed discussion of the D2AC module and procedure.

intended to do lexical transfer and not fully-fledged machine translation.

Table 1: Comparison of language specific components

	D2AC$_{2009}$	**D2AC$_{2010}$**
`Lex.LangIn-LangOut.txt`	2 696	2 740
`Lex.LangOut.txt`	350 943	385 599
`MorphRules.txt`	94	80
`G2GRules.txt`	53	57

A human assessment of the translation outputs shows that the output of the two systems contains similar errors: both leave a large number of words untranslated and have difficulty translating specific syntactic constructions (such as the negative construction) correctly. In addition, GT seems to translate Dutch compounds ineffectively, since compounds are consistently split into constituents, before being translated (e.g. Du. *vervoersituaties* is incorrectly translated to **vervoer situasies* instead of the correct Afr. *vervoersituasies* 'transport situations').

Table 2: BLEU scores of D2AC2010 and GT

	D2AC$_{2010}$	**GT**
% of 1-gram matches	58.88	72.32
% of 2-gram matches	29.36	46.21
% of 3-gram matches	16.76	33.09
% of 4-gram matches	9.59	23.57
BLEU	0.22	0.40

III. An Afrikaans-to-Dutch Convertor

A. Development of A2DC

Given the relative success of D2AC, we also wanted to experiment with the conversion of Afrikaans to Dutch. Since A2DC would have to do the exact opposite of D2AC, the relevant language-pair specific components of D2AC$_{2010}$ (i.e. Lex.LangIn-LangOut.txt, G2GRules.txt and MorphRules.txt) were reversed semi-automatically to create a base-line system from which A2DC could be developed. After the reversal, the rules contained in G2GRules.txt and MorphRules.txt were re-ordered to ensure that more specific rules would be executed first. Also, very general rules, such as the rule stating that any word can be suffixed with a *d* or *t*, were removed and replaced with more specific rules where possible.

Since it is sometimes not easy to decide whether a rule should be included in MorphRules.txt or in G2GRules.txt (e.g. the rule stating that the Afrikaans word ending **-sie** should be replaced by **-tie** to convert Afr. *petisie* to Du. *petitie* 'petition'), some experiments were done in this regard during a manual rule-order optimisation process. "Ambiguous" rules (i.e. rules which are not clear-cut morphological or graphological rules) were divided in a way which seemingly optimised the performance of the convertor. However, further investigation is needed to ensure an optimal rule division and automatic rule-induction will also be used to ensure the optimality of the rule-ordering and division. The resulting list of A2DC's MorphRules.txt contains 62 entries, while G2GRules.txt contains 80 entries.

The Lex.LangIn-LangOut.txt used in D2AC was also reversed so that it contained a list of Afrikaans entries (one entry per line) with possible Dutch translation alternatives. To enrich the Lex.LangIn-LangOut.txt a bilingual list of function words was developed and added to the existing entries in the lexicon. The resulting Lex.LangIn-LangOut.txt used in A2DC contains 2 474 entries. This list is smaller than the Lex.LangIn-LangOut.txt of D2AC$_{2010}$ due to the fact that different Dutch lexemes were translated to the same Afrikaans lexeme (e.g. Du. *attachment* and *bijlage,* which both translate to Afr. *aanhegsel* 'attachment'). Therefore, when the list was reversed, multiple Dutch entries were combined into one Afrikaans entry (i.e. *aanhegsel attachment*//*bijlage*). The lemma list of e-lex [4] was used as a Dutch lexicon (i.e. as Lex.LangOut.txt) and contains 412,683 Dutch entries.

B. A2DC as part of technology recycling

Given the fact that A2DC was not developed as a fully-fledged machine translation system but as a lexical convertor to aid technology transfer, we used it as part of a process of technology recycling. In this experiment we investigated the efficiency of A2DC as a pre-processing step for the recycling of a Dutch part of speech tagger. To evaluate this, Tadpole, a morphosyntactic tagger and parser for Dutch [5], was used to tag an Afrikaans translation of the METIS II test data as a first phase in the experiment. In the second phase, the same Afrikaans data was converted with A2DC before being tagged with Tadpole. The METIS II test data consists of 200 sentences from the Parole Corpus (see http://korpus.dsl.dk/e-resurser/parole-korpus.php). The Afrikaans translation was done by a professional translator and the resulting Afrikaans data set consists of 1 963 Afrikaans words.

Tadpole uses the CGN tagset [6], which consists of 12 part of speech categories, each of which is further divided using various tag specifications, resulting in a very large tagset consisting of more than 300 different tags. Since many of the tag specifications explicated in the CGN tagset are not applicable to Afrikaans (e.g. gender specification of common nouns), and given the fact that this evaluation is done on a small test set (200 sentences), only the twelve main tag categories were used in this experiment. Besides POS tags, the Tadpole output also contains lemmas and morphological analyses of annotated words, but those features were ignored for the purpose of this experiment. The two phases of this experiment will be discussed in the following sections.

IV. Results

A. Word-level evaluation

For the word-level evaluation, 500 words were randomly extracted from the 5,000 most frequent words in the Spoken Dutch Corpus [7][2]. Capitalised words (e.g. proper names and acronyms), abbreviations, and interjections were removed and

[2] This is the same dataset that was used for the word-level evaluation of D2AC in [2].

replaced with other randomly-selected words from the CGN frequency list. The resulting wordlist was manually translated to Afrikaans. Where more than one Afrikaans translation alternative existed, the most probable translation alternative was selected, resulting in a list of 500 Afrikaans words.

The resulting Afrikaans list was then translated back to Dutch, taking care to ensure that all possible Dutch translations were added. A2DC was then used to translate the 500 Afrikaans words and the output was compared to the manual Dutch translations. The results of this evaluation are given in Table 3.

Table 3: Results of word-level evaluation

	# tags assigned	# tags correct
<Lex.LangIn-LangOut>	92	92
<Lex.LangOut>	199	199
<Translated>	77	71
<Untranslated>	132	0
TOTAL	**500**	**362**

A2DC was able to provide at least one Dutch translation alternative for 72.4% (i.e. 362 words) of the words in the evaluation test set. Although the conversion modules only translated 77 words, 71 of the attempted translations were correct. Incorrect translation attempts include Afr. *uitgawe* 'expense' translated to Du. *uitgaven* 'expenses'. This means that the conversion modules have a conversion precision of 92.21%. The high precision is, once again, due to the fact that translated words are looked up in Lex.LangOut.txt to ensure that the word resulting from conversion is a valid Dutch word.

Table 4: Error analysis of word-level evaluation

Cause	# of untranslated words	% of untranslated words
Not in Lex.LangOut	5	3.79%
Not in Lex.LangIn-LangOut	18	13.64%
Not in Rules	20	15.15%
Past Tense Verbs	21	15.91%
Rule-ordering	68	51.51%
TOTAL	**132**	**100**

An analysis of the untranslated words shows that four factors are responsible for the bulk of the words not being converted (see Table 4). The first of these is the coverage of Lex.LangIn-LangOut.txt, since almost 14% of the words left untranslated are non-cognates or false friends and therefore should have been included in this lexicon. The Afr. *dieselfde*, for example, should be translated to Du. *hetzelfde*, but was not included in Lex.LangIn-LangOut.txt.

The second problem is the comprehensiveness of the rules included in MorphRules.txt and G2GRules.txt. More than 15% of the words left untranslated are instances that could have been handled by rules, and which should have been included in the system. So, for instance, Afr. *militêre* 'military' should have been translated to Du. *militaire*, but A2DC only contains a rule that converts *–êr* at the end of a word to *–air*, and the convertor is therefore unable to handle *–êre* correctly.

The conjugation of verbs is one major difference between Afrikaans and Dutch, since Afrikaans has a much simpler verb morphology. Therefore it is not surprising that the third hampering factor concerns the fact that A2DC is not able to handle the translation of Afrikaans past tense verbs effectively (e.g. Afr. *geken* 'knew' which should have been translated to Du. *kende/kenden/gekend*). It would, however, only be possible to successfully convert Afrikaans verbs to Dutch if some contextual syntactic information could be taken into consideration, but syntactic conversion currently falls without the scope of this research project.

The bulk of the untranslated words (51.51%) should have been converted by rules currently included in the system. So, for example, Afr. *beset* 'occupied' should have been translated to Du. *bezet* by the same rule that translated Afr. *versoek* 'request' into Du. *verzoek*, but *beset* was left untranslated while *versoek* was translated correctly. This is due to the ordering of the rules, and optimising A2DC would therefore entail a thorough reconsideration of the rule-ordering. Since the focus of our current research is on expediting the development of resources, we decided to not spend much more time on such optimization, in order to determine the efficiency of technology transfer even with less-than-perfect technologies.

B. Sentence level evaluation

Despite the fact that our current research is not aimed at providing a fully-fledged Afrikaans-to-Dutch machine translation system, we did an experiment to get an impression of how A2DC compares to another available solution, the Afrikaans-Dutch GT, when translating sentences. To do this comparative evaluation, we used the development test set that was used to evaluate the Dutch-English machine translation system in the METIS II project [8]. The Dutch METIS II sentences were translated to Afrikaans, and these Afrikaans sentences were sent to different translators to prepare reference translations with which the BLEU scores [9] could be calculated. The results for A2DC and GT are shown in Table 5.

Table 5: Results of sentence level evaluation

	A2DC	GT
% of 1-gram matches	53.54	73.69
% of 2-gram matches	22.04	49.75
% of 3-gram matches	10.63	36.95
% of 4-gram matches	5.29	28.15
BLEU	0.16	0.44

In this evaluation GT significantly outperforms A2DC as was expected, since A2DC is not able to handle any syntactic differences between Afrikaans and Dutch. This inability of A2DC is particularly apparent when one considers the percentage of n-gram matches shown in Table 5. While GT has a 4-gram match of almost 28%, only 5.29% of 4-grams in the A2DC output also occurred in at least one of the reference translations. The fact that A2DC only obtains a BLEU score of 0.16, while D2AC scored 0.22 in a similar experiment is also

not surprising, given that Dutch has a higher level of morphological complexity than Afrikaans.

A human assessment of the translations shows that, apart from the fact that A2DC leaves a large number of words (many of which are past tense forms of verbs) untranslated, it is also unable to handle syntactic issues like the Afrikaans double negation. Surprisingly, GT can also not handle the Afrikaans negation construction correctly and consistently retains the second negation particle (the second *nie* in negated Afrikaans sentences) after translating it to Du. *niet*. However, given the fact that A2DC was not developed as a fully-fledged machine translation system but as a lexical convertor to aid technology recycling, this evaluation does not give a good indication of the efficiency of the system. Therefore we also evaluated A2DC as part of a process of technology recycling, by using a Dutch part of speech (POS) tagger to annotate Afrikaans data. This evaluation experiment is described in the following section.

Table 6: Results per POS category (raw and converted Afrikaans data)

	Results on raw Afrikaans data			**Results on converted Afrikaans data**		
	Precision	Recall	*f*-score	Precision	Recall	*f*-score
N[3]	0.54	0.86	0.67	0.67	0.91	0.77
ADJ	0.61	0.73	0.66	0.64	0.78	0.7
V	0.86	0.61	0.71	0.89	0.62	0.73
NUM	1	0.79	0.88	0.97	0.76	0.86
PRON	0.34	0.55	0.42	0.84	0.88	0.86
ART	0.16	0.01	0.02	0.95	1	0.97
PREP	1	0.81	0.9	0.99	0.99	0.99
CONJ	0.65	0.59	0.62	0.96	0.86	0.91
ADV	0.64	0.85	0.73	0.78	0.7	0.74
INTERJ	0	0	0	0	0	0
SPEC	0.43	0.74	0.54	0.2	0.07	0.1

Table 7: Confusion matrix (raw Afrikaans data)

	N	**ADJ**	**V**	**NUM**	**PRON**	**ART**	**PREP**	**CONJ**	**ADV**	**INTERJ**	**SPEC**
N	**17.42**	1.27	4.02	0.31	1.32	2.75	1.63	0.71	1.68	0	0.82
ADJ	0.97	**6.16**	1.78	0	0.10	0	0.31	0	0.76	0	0.10
V	1.27	0.36	**12.12**	0	0.05	0.05	0.05	0.05	0.05	0	0.05
NUM	0	0	0	**1.32**	0	0	0	0	0	0	0
PRON	0	0.10	0.66	0.20	**5.40**	9.16	0	0.36	0.10	0	0
ART	0	0	0.82	0	0	**0.15**	0	0	0	0	0
PREP	0	0	0	0	0	0	**10.29**	0	0	0	0
CONJ	0	0	0	0	0.92	0	0	**1.73**	0	0	0
ADV	0	0.10	0.05	0.05	0.66	0	0.05	0	**5.20**	0	0
INTERJ	0	0	0	0	0	0	0	0	0	**0**	0
SPEC	0.71	0.41	0.41	0	1.27	0.05	0.36	0.10	0.36	0	**2.80**

Table 8: Confusion matrix (converted Afrikaans data)

	N	**ADJ**	**V**	**NUM**	**PRON**	**ART**	**PREP**	**CONJ**	**ADV**	**INTERJ**	**SPEC**
N	**18.54**	1.02	3.36	0.10	0.10	0	0.05	0	1.27	0	3.06
ADJ	0.31	**6.52**	2.09	0	0	0	0.05	0	0.82	0	0.36
V	0.92	0.46	**12.38**	0	0	0	0	0	0.10	0	0.10
NUM	0.05	0	0	**1.43**	0	0	0	0	0	0	0
PRON	0	0.10	1.07	0.25	**8.61**	0.05	0	0.10	0.10	0	0
ART	0	0	0.46	0	0.20	**12.07**	0	0	0	0	0
PREP	0	0	0	0	0	0	**12.53**	0.10	0	0	0
CONJ	0	0	0.10	0	0	0	0	**2.55**	0	0	0
ADV	0.05	0.15	0.20	0.10	0.87	0	0.05	0.20	**5.70**	0	0
INTERJ	0	0	0	0	0	0	0	0	0	**0**	0
SPEC	0.51	0.15	0.20	0	0	0	0	0	0.15	0	**0.25**

3 In tables 6 – 8 and in Section C the following tag abbreviations are used: noun (N), adjective (ADJ), verb (V), numeral (NUM), pronoun (PRON), article (ART), preposition (PREP), conjunction (CONJ), adverb (ADV), interjection (INTERJ) and special tokens (SPEC).

C. A2DC as part of technology recycling

1) Tagging raw Afrikaans data with a Dutch POS tagger

In the first phase of the experiment, the complete Afrikaans test set was tagged with Tadpole. Tag specifications, lemmas and morphological analyses were removed from the Tadpole output, and clustered words were manually separated (e.g. *onder_andere VZ_ADJ* were separated into *onder VZ andere ADJ*). The POS annotations provided by Tadpole were then manually checked and corrected to create a Gold Standard. The Tadpole output was compared to the Gold Standard and accuracy, recall, precision and *f*-scores were calculated for each POS category. These results are given in Table 6 while

Table 7 shows a confusion matrix with relative values for the POS categories[4]. For the calculations, punctuation was not included.

The all-over accuracy of the Tadpole output on unconverted Afrikaans data is 62.6%. Most POS categories have a rather low *f*-score (see Table 6), except for NUM and PREP, which have *f*-scores of 0.88 and 0.90 respectively. The confusion matrix (Table 7) also indicates that instances of NUM and PRON were very seldom mistagged. The good results for these categories are due to the fact that many Afrikaans and Dutch numerals are identical in form (e.g. *eerste* 'first', *twintig* 'twenty', *op* 'on', *onder* 'below', etc.). In contrast, ART have a very low f-score (0.02), because the Afrikaans definite article *die* 'the' is an unambiguous pronoun in Dutch. The very frequently occurring Afrikaans *die* is therefore consistently tagged as a pronoun, resulting in the high ART-PRON confusion (see

Table 7). N, ADJ, V, PRON, CONJ, ADV and SPEC have relatively average *f*-scores, ranging from 0.42 to 0.73.

Tadpole uses N as default tag and therefore N is the tag that is assigned to all instances that Tadpole is unable to disambiguate correctly. This leads to a relatively high rate of N confusion for all categories (see Table 7). This is especially true in the case of verbs. Because Afrikaans verbs are not conjugated in the same way as Dutch verbs, many verbs are not in the correct form, given the context in which they occur. This makes it impossible for Tadpole to determine that a word should be assigned the tag V and Tadpole then assumes that the word is a noun and should be tagged with N. This results in a high V-N confusion.

A manual assessment of the data showed that, apart from the definite article, Tadpole also persistently tags the Afrikaans verb *het* 'have' (Du. *hebben*) incorrectly. This is because *het* can only be a definite article or a pronoun (but never a verb) in Dutch. Furthermore, Tadpole erroneously tags the Afrikaans first person personal pronoun *ek* 'I' (Du. *ik*) as a conjunction and the Afrikaans indefinite article *'n* 'a' (Du. *een*) as a noun. The fact that function words, such as PRON, ART, PREP and CONJ are often erroneously tagged is problematic, since these classes are used as anchors for the rest of the tagging task. These categories are closed classes and therefore Tadpole assigns the tags for these classes with a high level of certainty. They then serve as contextual information to assign the tags of open classes (N, V, ADJ and ADV). The erroneous tagging of these words leads to incorrect contextual information which, in turn, has a negative effect on the accuracy of the tagger on open classes.

2) Tagging converted Afrikaans data with a Dutch POS tagger

The Afrikaans translation of the METIS II test set was once again used as data in this second phase of the experiment. This time, instead of sending raw Afrikaans data through Tadpole (as was done in the first phase described in the previous section), we first converted the Afrikaans test set with A2DC to make the data more "Dutch-like". The basic idea here is to eliminate errors such as those described above by first converting the Afrikaans text to Dutch. Even though this conversion will not yield a good Dutch translation, we hypothesise that a Dutch POS tagger will benefit from the conversion, since frequently occurring false friends and non-cognates could be handled better (or even correctly) after they had been converted to Dutch. After conversion the A2DC output was tagged with Tadpole, and the Tadpole output was once again compared to the Gold Standard. Precision, recall and *f*-scores are shown in Table 6, while Table 8 shows the confusion matrix.

The accuracy for the A2DC Tadpole output is 80.6%, which is a considerable improvement over the 62.6% obtained in the first phase of this experiment, and which clearly illustrates the value of using conversion as a pre-processing step for technology recycling. The confusion matrix shows that the accuracy for all classes, except for SPEC, increased. The *f*-scores (see Table 6) for PRON and especially for ART are much higher than in the experiment with the raw Afrikaans data. A2DC successfully converted Afrikaans articles into their Dutch equivalents (i.e. A2DC translated Afr. *'n* to Du. *een* 'a' and Afr. *die* to Du. *de* or *het* 'the'), causing the *f*-score for ART to increase from 0.02 to 0.97.

Other POS categories that show an evident *f*-score improvement are PRON (from 0.42 to 0.86), CONJ (from 0.62 to 0.91) and PREP (from 0.9 to 0.99). In the case of N, ADJ, ADV and V a slight gain is noticeable. The score for numerals is almost the same as in the first part of the experiment. The only category that shows a decrease in *f*-score (from 0.54 to 0.10) is the category of special tokens. The words that receive tag SPEC are mainly parts of multiword proper nouns, e.g. Afr. and Du. *Europese SPEC Unie SPEC* 'European SPEC Union SPEC'. Many such instances in the converted data are tagged incorrectly, since A2DC changes all words into lower case, which makes it hard for Tadpole to recognize these words as proper nouns. Given the lower case version, Tadpole tags instances like these as *europese ADJ unie N*.

Table 8 reiterates the improvements over all categories except SPEC. It is apparent though that a large number of verbs are still erroneously tagged as nouns. This is because A2DC is unable to convert verbs into the correct form, and therefore these verbs are still assigned the default tag, N. Other

[4] Rows indicate the Tadpole output; columns refer to the true classes.

categories are consistently confused with N less often and ART and CONJ are never mistakenly tagged as N after conversion with A2DC. This is because articles and conjunctions are closed classes which are correctly converted by the bilingual lexicon included in A2DC. The fact that closed classes are assigned the correct tag more often after conversion, improves the tagger's ability to correctly assign tags to the open classes.

The fact that it was possible to develop an Afrikaans POS tagger achieving an accuracy of more than 80% using the technology recycling approach, is especially noteworthy when it is compared to the Afrikaans POS tagger development described in [10]. In this project an Afrikaans POS tagger was developed using the TnT tagger 0. Given the fact that no annotated Afrikaans data was available at the time, the development of the POS tagger necessitated the manual annotation of data to train the TnT algorithm. The Afrikaans TnT tagger, using 13 tags, was only able to achieve an accuracy of more than 80% after the manual annotation of 1 999 words. The use of A2DC and technology recycling could therefore have expedited the development described in [10], since it would not have been necessary to manually tag the first batch of training data. Of course, to have developed A2DC also required some effort and time, but now that a conversion module is available in the open source domain (sourceforge.net/projects/d2ac-a2dc), it could benefit the development of technologies for Afrikaans that already exist for Dutch (e.g. semantic parsers, chunkers, etc.).

V. Conclusion and Future work

In this article we described the development and performance of an A2DC, which achieved an accuracy of 72.4% in a word-level evaluation and a BLEU score of 0.16 in a sentence level evaluation. A2DC was then used as part of a technology recycling process and the use of the convertor improved part of speech tagging accuracy by 18% (from 62.6% without conversion to 80.6% when using conversion). In the evaluation experiments we found that handling of especially verb conjugations, false friends and non-cognates require close attention. Most of these issues could be addressed by refining and/or extending the bilingual translation list (Lex.LangIn-LangOut.txt) (semi-) automatically – for instance by using the CELEX database to complete verb paradigms and plural forms of nouns, and by using semi-automatic processes to extend the list of false friends. In addition, some more attention should be paid to rule-ordering addition, as well as iteration of rules and the effect thereof on the greediness of the conversion modules. Automatic rule-induction should also be investigated.

An important conclusion of this paper is that rule-based conversion could play a major role in expediting technology development for resource-scarce languages, specifically when technology recycling is used. Further research needs to be done in order to determine whether this observation holds true for tasks other than part of speech tagging. Work in the near future will therefore include using A2DC in technology recycling tasks involving other resources, such as an Afrikaans lemmatiser, chunker and morphological analyser. The approach used here could, in principle, also be extended to other resource-scarce languages. Given the fact that all the indigenous South African languages are considered resource scarce, future work will include the use of the approach described in this paper as a way in which technology development for these languages can be expedited.

VI. Acknowledgements

Part of this research was made possible through a research grant by the South African National Research Foundation (FA207041600015). We would like to extend our gratitude to the following people who were involved in various aspects of the project: Kirsten Arnauts, Veronique de Gres, Shanna Pettens, Martin Puttkammer, Carla-Mari van den Heever, and Daan Wissing. All fallacies remain ours.

References

[1] Rayner, M., Carter, D., Bretan, I., Eklund, R., Wiren, M., Hansen, S.L., Kirchmeier-Andersen, S., Philp, C., Sorensen, F., and Thomsen, H.E., "Recycling lingware in a multilingual MT system", In Burstein, J., and Leacock, C. (eds.), *From research to commercial applications: making NLP work in practice*, ACL, Somerset, 1997, pp 65-70.

[2] Van Huyssteen, G.B. & Pilon, S., "Rule-based conversion of closely-related languages: A Dutch to Afrikaans convertor", *Proc. of the 20th Annual Symposium of the PRASA,* Stellenbosch, South Africa, 2009, pp 23-28.

[3] Van Huyssteen, G.B. & Pilon, S, "A Dutch-to-Afrikaans convertor", *Poster presented at the 20th CLIN Meeting*, Utrecht, The Netherlands, 2010.

[4] Dutch Language Union (Nederlandse Taalunie, NTU). 2006. *E-Lex Version 1.1*. Institute for Dutch Lexicology, Leiden, The Netherlands.

[5] Van den Bosch, A., Busser, G.J., Daelemans, W., and Canisius, S., "An efficient memory-based morphosyntactic tagger and parser for Dutch", F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (eds.), *Selected Papers of the 17th CLIN Meeting*, Leuven, Belgium, 2007, pp 99-114.

[6] Van Eynde, F., Zavrel, J. & Daelemans, W., "Part of speech tagging and lemmatisation for the Spoken Dutch Corpus", In M. Gavrilidou et al. (eds.), *Proc. of the 2nd Intern. LREC*, Athens, Greece, 2010, pp 1427-1433.

[7] Nederlandse Taalunie, *Corpus gesproken Nederlands 1.0*, TST-Central, Leiden, 2004, [Web:] http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm [Accessed on 2009/09/26].

[8] Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M., and Schmidt, P., "Evaluation of a machine translation system for low resource languages: METIS-II", *Proc. of the 6th intern. LREC*, Marrakech, 2008, pp 449-456.

[9] Papineni, K., Roukos, S., Ward, T., and Zhu, W.J., "BLEU: a method for automatic evaluation of machine translation", *Proc. of the 40th an. meeting of the ACL*, Philadelphia, 2002, pp 311-318.

[10] Pilon, S. 2005. Outomatiese Afrikaanse woordsoortetikettering. North-West University (Potchefstroom Campus). Potchefstroom. (Dissertation – MA).

Brants, T., "TnT – A statistical part-of-speech tagger", *Proc.of the Sixth Applied Natural Language Processing Conference,* 2000, [Web:] http://acl.ldc.upenn.edu/A/A00/A00-1031.pdf [Date accessed: 2004-03-03]