# A Rule-based Shallow-transfer Machine Translation System for Scots and English

**Gavin Abercrombie**
jst662@alumni.ku.dk

**Abstract**

An open-source rule-based machine translation system is developed for Scots, a low-resourced minor language closely related to English and spoken in Scotland and Ireland. By concentrating on translation for assimilation (gist comprehension) from Scots to English, it is proposed that the development of dictionaries designed to be used within the `Apertium` platform will be sufficient to produce translations that improve non-Scots speakers understanding of the language. Mono- and bilingual Scots dictionaries are constructed using lexical items gathered from a variety of resources across several domains. Although the primary goal of this project is translation for gisting, the system is evaluated for both assimilation and dissemination (publication-ready translations). A variety of evaluation methods are used, including a cloze test undertaken by human volunteers. While evaluation results are comparable to, and in some cases superior to, those of other language pairs within the `Apertium` platform, room for improvement is identified in several areas of the system.

**Keywords:** Scots, machine translation, assimilation

## 1. Introduction

The Scots language is spoken by over 1.5 million people in Scotland[1] and a further 140,000 in Ireland,[2] but is low-resourced in terms of technology - as far as I am aware there are currently no automatic translators available for this language. The development of machine translation (MT) systems can lead to positive outcomes for a minor language and its speakers by contributing to the language's standardisation, and assisting the diffusion of content originally created in that language (Forcada, 2006). The aim of this project is to take a step towards the 'de-minorizing' of Scots by developing a shallow transfer, rule-based MT system as part of the `Apertium` platform.

Since all Scots speakers comprehend standard English,[3] this study will focus on translation from Scots to English for assimilation (or gisting) purposes. With assimilation thought to be the most common use of MT systems today (O'Regan et al., 2013), this would seem to be an effective use of time and resources as an initial goal in MT between these two languages. As the two languages are very closely related, it is proposed that translation of the bulk of lexical items in Scots texts (combined with part-of-speech (POS) analysis) will be sufficent to enable gist comprehension by non-Scots speakers.

## 2. Background

### 2.1. The Scots language

Scots is a West Germanic language descended from Old Northumbrian (a branch of Old English) and closely related to modern English (Macafee and Aitken, 2002). Indeed, it shares many characteristics with its neighbour and relative including SVO sentence structure and a 'vast body' of common vocabulary (Tulloch, 1997).

Grammatical differences can however be seen in the word order of some constructions e.g. *whit fur did he dae that?* ('what did he do that for'). There are also notable differences to English in negation e.g. *huvnae* ('have not'), *dinnae* ('do not'), and the double modal e.g. *he'll can dae that* ('he'll be able to do that') (Beal, 1997).

While Scots shares a great deal of vocabulary with English, albeit often with different forms and/or senses, it also includes a large quantity of words derived from Old English that are not present in English, as well as loan words from Gaelic, French and Latin, the Scandinavian languages, and other West Germanic languages such as Frisian (Jones, 1997).

There is no official standard written form of Scots,[4] so spelling varies greatly - any translation system needs to account for multiple forms of many, if not most, words.

### 2.2. Translation: assimilation and dissemination

The aims of translation between languages can vary. *Dissemination* is 'the production of translations of 'publishable quality' (Hutchins, 2003a), and can help in the creation of more text in a lesser-resourced language,[5] while *assimilation* is translation for gist comprehension by readers who 'can accept poor quality as long as they can get an idea of what the text conveys' (Hutchins, 2003b).

Successful assimilation is often viewed as the more realistic and achievable goal for MT systems (Hutchins, 2003a) and is probably 'the most frequent application of MT nowadays' (O'Regan et al., 2013). Translation for assimilation would enable people who do not speak the language to understand Scots text 'thus removing an argument against writing in the lesser-resourced language,'[6] and is the primary aim of this study.

### 2.3. Machine translation

Approaches to machine translation can be broadly categorized as either rule-based or statistical, or in some cases, a hybrid of the two.

[1] Scotland's Census 2011 - National Records of Scotland

[2] 2001 Census of Northern Ireland - Northern Ireland Statistics & Research Agency

[3] Scotland's Census 2001 - National Records of Scotland

[4] www.educationscotland.gov.uk/knowledgeoflanguage/scots/writinginscots/scotsspelling/index.asp

[5] wiki.apertium.org/wiki/Assimilation_and_Dissemination

[6] wiki.apertium.org/wiki/Assimilation_and_Dissemination

While statistical MT systems have the advantage of not requiring the explicit programming of language rules, they generally require large amounts of bilingual corpus data in order to produce satisfactory results. As Forcada (2006) suggests, lacking such resources, 'it may be much easier for speakers of the minor language to encode the language expertise needed to build a rule-based machine translation system.'

While statistical MT tends to produce more fluent translations, those of rule-based systems are often more faithful to the details of content of the original text (Forcada et al., 2011), a somewhat important advantage where information transferral is concerned.

Due to the low-resource status of Scots, and this study's focus on translation for assimilation, rule-based translation would currently seem to be the more appropriate approach.

#### 2.3.1. Shallow-transfer rule-based machine translation

Most rule-based MT systems create translations by parsing the source text, creating an intermediate symbolic representation of it, and then generating a final translation in the target language. They apply mappings between lexical items stored in dictionaries as well as transfer rules to account for structural differences between the two languages. Translation of unrelated or distantly related languages requires deep syntactic and semantic analysis, whereas closely related languages (such as English and Scots) can be translated with shallow parsing (Sánchez-Martínez and Forcada, 2009).

#### 2.3.2. The `Apertium` MT platform

`Apertium` is a free, open-source platform for developing rule-based, shallow-transfer machine translation systems, which was initially developed for translation between closely related languages (Forcada et al., 2011). As 'one of the few open-source MT systems that can be used for real-life purposes' (Forcada, 2006), the `Apertium` platform is highly suitable for this project.

The system consists of a number of modules through which the input text is passed and modified before a translation in the target language is outputted. Figure 1 illustrates this pipeline.

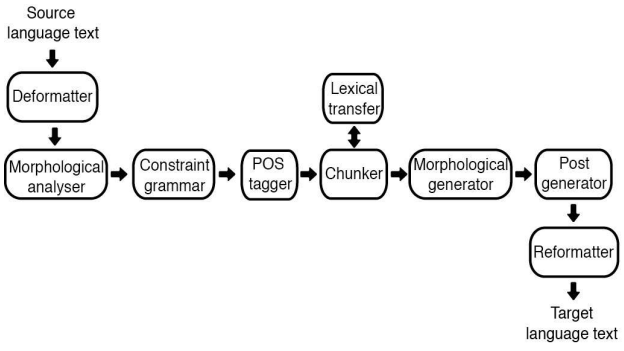


Figure 1: Architecture of the `Apertium` system. Adapted from Forcada et al., (2011).

The modules perform the following functions:

- The *deformatter* and *reformatter* modules handle formatting and blank spaces.
- The *morphological analyser* maps each word (or multi-word unit) in the input to entries in a monolingual dictionary, returning their lemma, lexical category and morphological information. This information is passed to the other modules and eventually mapped to a target language surface form by the *morphological generator*.
- A *constraint grammar* module contains rules designed to reduce ambiguity regarding parts-of-speech.
- The *part-of-speech tagger* selects the most statistically likely of the possible lexical forms, given an item's context.
- The *lexical transfer* module maps the given lexical item to its corrsponding lemma in the bilingual dictionary.
- The *chunker* segments the input into syntactic chunks and handles transfer rules.
- The *post generator* handles orthographic operations such as contractions.

Further details of the `Apertium` architecture can be found in Forcada et al. (2011).

## 3. Development of the English-Scots translation system

In the belief that translation of Scots-only lexical items to English will be sufficient to enable gist comprehension, this study concentrates on the development of the monolingual and bilingual dictionaries for this language pair, and particularly on including Scots words that do not exist in English. Development of the system was undertaken in the following stages.

### 3.1. Prior status of the `eng-sco` language pair in the `Apertium` platform

`Apertium` language pairs are classified as being in one of four stages of development: from the *trunk* for release quality pairs, down through *staging* and *nursery* to *incubator* where incomplete 'dictionaries, dictionary fragments, rules, things that aren't quite ready to live in the real world' are kept.[7] Prior to this study, English & Scots (`eng-sco`) was in this latter category, and contained a Scots dictionary of only 172 entries.

### 3.2. Construction of `eng-sco` translator modules

An English monolingual dictionary containing approximately 20,000 lexical entries and corresponding morphological information was first added to `eng-sco` in the `Apertium` *incubator*. Associated dictionary processing files such as the *post generator* `apertium-eng.post-eng.dix` were also included.
Work then began on expanding the Scots dictionaries.

[7] http://wiki.apertium.org/wiki/Incubator

### 3.3. Acquisition of linguistic data

Scots lexical items were collected from the following sources and added to the monolingual Scots dictionary, the English-Scots bilingual dictionary, and where necessary the monolingual English dictionary.

- The *Wiktionary* of Scots words[8] containing 371 entries (some of which are different forms of the same lemma, such as *airts*, *airtin*, *airit*, 'direct' or 'guide') was scraped and these items were added.
- A number of Scots texts were manually examined, adding any previously absent words to the dictionaries. Efforts were made to cover a broad range of domains by using a wide variety of document types such as fiction, government documents, and social media content. This yielded some 2,500 lexical items.
- In order to gain maximum lexical coverage according to Zipf's law (Manning and Schütze, 1999), the dictionaries were now checked against a list of the most common English words.[9] Where these, or their Scots translations, were not present, they were added to the Scots and bilingual dictionaries.
- Finally, the process was begun of adding items present in the English monolingual dictionary, and their Scots translations to the Scots and bilingual dictionaries.

### 3.4. Dictionary construction

In the monolingual dictionaries, entries consist of a *lemma*, an *identity* - the word's root form, and a *paradigm*, which refers to information about how the word should inflect. In some cases, e.g., alternate spellings, a transfer direction rule is added, to ensure that only a preferred variant of an entry is translated in one direction or the other.
Entries containing a Scots lemma along with its English meaning are added to the bilingual dictionary, and the English monolingual dictionary is checked to ensure the presence of the word's translation.

Following these steps, at the current stage of development the bilingual and Scots dictionaries contain around 3,000 individual lemmas (plus 5,758 personal names).

### 3.5. Translation from Scots to English with the `eng-sco` system

The following illustrates how an input sentence is modified as it passes through the `Apertium eng-sco` translation system, and the resulting output at each stage. (Individual words are preceded by a ∧ symbol and followed with a $).

**Input**

*Ye can play hunners o tricks.*

The above example sentence is input to the MT system.

↓

**Morphological analysis**

$$\wedge Ye/Prpers < prn >< subj >< p2 >< mf >< sp > /Prpers < prn >< obj >< p2 >< mf >< sp > \$ \wedge can/can < n >< sg > /can < vaux >< pres > \$ \wedge play/play < vblex >< inf > /play < vblex >< pres > \$ \wedge hunners/hunner < n >< pl > \$ \wedge o/o < pr > \$ \wedge tricks/trick < vblex >< pri >< p3 >< sg > \$.$$

Each word is looked up in the Scots dictionary, and if present, its lemma and morphological information are returned e.g., *hunners* is identified as being the plural form (*pl*) of the noun (*n*) *hunner*. In another example, as *trick* is only present in the dictionary as a verb (and not as a noun), the third-person (*p3*) singular (*sg*) present indicative (*pri*) form of the lexical verb (*vblex*) is returned.

↓

**Part-of-speech tagging**

$$\wedge Prpers < prn >< subj >< p2 >< mf >< sp > \$ \wedge can < vaux >< pres > \$ \wedge play < vblex >< inf > \$ \wedge hunner < n >< pl > \$ \wedge o < pr > \$ \wedge trick < vblex >< pri >< p3 >< sg > \$.$$

Here the tagger chooses the most likely part-of-speech tag for each word given the sequence of words in the sentence. Where two or more options are available, the other possibilites are discarded. In the case of *Ye* for example, the tagger assesses it to be a subject pronoun (*prn*, *subj*, rejecting the possibilty generated in the previous module of it being an object (*obj*) pronoun.

↓

**Lexical transfer**

$$\wedge Prpers < prn >< subj >< p2 >< mf >< sp > /Prpers < prn >< subj >< p2 >< mf >< sp > \$ \wedge can < vaux >< pres > /can < vaux >< pres > \$ \wedge play < vblex >< inf > /play < vblex >< inf > \$ \wedge hunner < n >< pl > /hundred < n >< pl > \$ \wedge o < pr > /of < pr > \$ \wedge trick < vblex >< pri >< p3 >< sg > /trick < vblex >< pri >< p3 >< sg > \$.$$

These items are now searched for in the billingual dictionary and the lemmas of the Scots and English entries are outputted e.g., *o* and *of*, *can* (Scots) and *can* (English).

↓

**Chunking / transfer rules**

$$\wedge Prpers < prn >< subj >< p2 >< mf >< sp > \$ \wedge can < vaux >< pres > \$ \wedge play < vblex >< inf > \$ \wedge hundred < n >< pl > \$ \wedge of < pr > \$ \wedge trick < vblex >< pri >< p3 >< sg > \$.$$

The corresponding items are found in the English dictionary. Syntactic chunking and transfer rules are also applied here, although in this case none are applicable.

[8] en.wiktionary.org/wiki/Index:Scots

[9] en.wikipedia.org/wiki/Most_common_words_in_English#Parts_of_speech

↓

**Morphological generation**

*You can play hundreds of tricks.*

According to the received morphological information, this module outputs the corresponding form of each word from the English dictionary.

↓

**MT output**

*You can play hundreds of tricks.*

These then form the final output produced by the MT system. In this case, despite the incorrect evaluation of *trick* at the morphological analysis stage, a correct[10] translation has been produced.

## 4. Evaluation

Several different evaluation methods are used to assess the effectiveness of the MT system. Evaluation was performed on texts taken at random from the Scottish Corpus of Texts & Speech,[11] which had not been used in development of the MT system. Reference translations were created manually by this author. [12]

### 4.1. Error Rate: WER and PER

Although the focus of this project is on translation for assimilation rather than dissemination, it is interesting to see how far away the system is from producing publication-ready translations. Word Error Rate (WER) compares the source text, translation, and a reference translation, calculating the minimum number of errors (*insertions* (*I*), *deletions* (*D*), and *substitutions* (*S*)) that require correction in order that the translated text matches the original (Koehn, 2009). Equation 1 shows the WER formula where *N* is the total number of words in the reference text.

$$WER = \frac{I + D + S}{N} \quad (1)$$

Position-independent Error Rate (PER) is a similar measure which takes into account the fact that translations with differing word orders can be equally valid (Koehn, 2009).

### 4.2. Common N-grams: BLEU

Measures that assess the performance of translators on N-grams that are common between reference and system translation are a popular way of tackling the problem of word order in MT (Koehn, 2009). The most popular of these is the `BLEU` algorithm (Koehn, 2009), which further addresses the problem of missing words by including a 'brevity penalty' (Lin and Och, 2004), and is used here to evaluate the `eng-sco` translator.

Error rate and common N-gram scores are calculated using `Asiya` (Gonzàlez et al., 2012)

### 4.3. Assimilation evaluation

There is no one established method for assimilation evaluation of MT (Ageeva et al., 2015), but cloze testing, as used by Trosterud and Unhammer (2012), O'Regan et al. (2013), and Ageeva et al. (2015), has the advantage of being less costly and subjective than methods that require billingual experts (O'Regan et al., 2013).

Evaluation proceeds as follows: volunteers with no prior knowledge of Scots are presented with sentences from which a number of words have been removed. Their task is to complete these gaps with an appropriate item selected from a list of candidates. Following O'Regan et al. (2013) and Ageeva et al. (2015), sentences are presented in the following conditions:

- Reference sentence only: a baseline score accounting for gaps that can be filled by guessing, e.g., due to common collocations.
- Reference sentence and source sentence: a further baseline assessing how much the original text assists comprehension, e.g., with proper nouns, loan words or close cognates.
- Reference sentence and MT sentence: this task assesses the contribution of the MT system to comprehension.
- Reference sentence, source sentence, and MT sentence: here the complimentary effect of the source and MT hints is assessed, as in a real-world translation situation.

Gaps are created by removing 30%[13] of words in each sentence from the following POS categories: *noun* (including *proper nouns*), *adjective*, *adverb* and *lexical verb*, which are deemed to be likely content words (following Ageeva et al. (2015)). The list of candidate words to be inserted comprises all the words from the same POS category as the correct missing item.

32 sentence sets were prepared for evaluation and these were divided into two blocks of 16 in order to limit the task length for volunteer human participants. The order of presentation of the four conditions (as described above) and that of the word options were randomized.

Each block of 16 sentence sets was evaluated by four volunteers. Participants were either native English speakers or highly proficient in English, but had no knowledge of Scots. They were aged between 24 and 34 and were all masters level students studying in English. The evaluation task was completed remotely online.[14]

[10] i.e., conforming to human translation. The input example is the first six words of the *The Eedjits* (Dahl, R. et al., Black and White 2006), the Scots translation of *The Twits* (Dahl, R., Cape 1980), and the original sentence does indeed read '*You can play hundreds of tricks*.'

[11] www.scottishcorpus.ac.uk

[12] Source, reference and MT sentences used for evaluation can be viewed at https://docs.google.com/document/d/1Q720RCbyPfVpemAua01HeF2uivJBJcp4OhDzy1Qtg1U/edit?usp=sharing

[13] Ageeva et al. (2015) experimented with varying this percentage, but achieved inconclusive results. It was felt that as Scots and English are so similar, a lower number of removed words may produce overly simple, easily guessed tasks.

[14] Tasks can be viewed at

| | |
|---|---|
| **Task** | Can you __1__ searching for __2__ and toads to __3__ in a __4__ sweetie __5__. |
| **Ref** | Can you remember searching for frogs and toads to keep in a big sweetie jar. |
| **Src** | Kin ye mind o searchin' fur puddocks and taddies tae keep in a big sweetie jar. |
| **MT** | Kin you mind of searchin' fur toads and taddies too keep in a big sweetie jar. |

Table 1: An example gapped task sentence with the three hint sentence types: reference, source and machine-translated.

## 5. Results

### 5.1. Dissemination results

Error rates and common N-gram scores were calculated on 72 sentences of between ten and 20 words taken at random from the *Scottish Corpus of Texts & Speech*. Overall scores can be seen in Table 2.

| Measure | Score |
|---|---|
| **WER** | 30.79 |
| **PER** | 29.16 |
| **BLEU** | 0.43 |

Table 2: Error rate and common N-gram scores for translation from Scots to English with `eng-sco`. Lower percentages are preferable for error-rate measures (WER and PER), while high scores are sought in the common N-gram evaluation (BLEU).

With error rates of around 30% for both WER and PER, the Scots-English translations compare reasonably well with published results[15] for released pairs (see Figure 2).

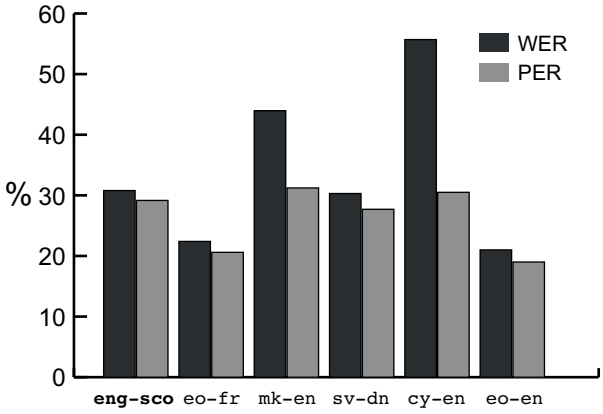


Figure 2: Comparison of WER and PER scores for `eng-sco` and other language pairs.

These scores are very similar to those of closely-related pair Swedish-Danish (`sv-dn`, WER: 30.3%; PER: 27.7%), and approaching those of other language pairs that are well established within the `Apertium` platform such as Esperanto-French (`eo-fr`, 22.4%; 20.6%) and Esperanto-English (`eo-en`, 21%; 19%). Error rates for `eng-sco` also compare favourably with less closely related pairs such as Macedonian-English (`mk-en`, 43.96%; 31.22%) and Welsh-English (`cy-en`, 55.7%; 30.5%), particularly for WER (for which the similarity in word order of English and Scots is an obvious advantage).

The translator achieves a `BLEU` score of 0.43, a result which lies between those published for Norwegian Nynorsk-Norwegian Bokmål (`nn-nb`: 0.74) and English-Esperanto (`en-es`: 0.1851).

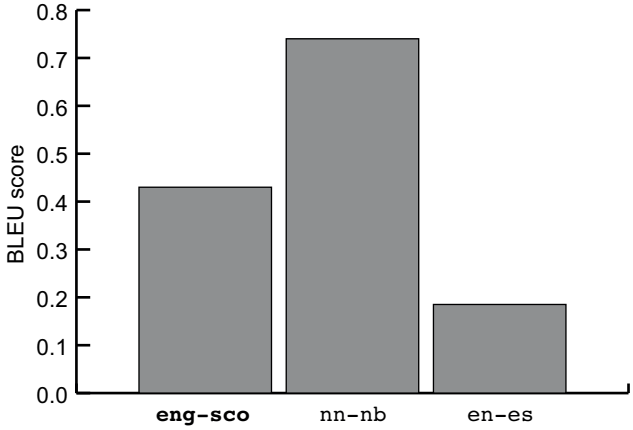


Figure 3: Comparison of BLEU scores for `eng-sco` and other language pairs.

### 5.2. Assimilation results

Averaged assimilation evaluation task results can be seen in Table 3. As expected, all three hint conditions score considerably higher than when no hint in available. Furthermore, the number of correct answers is higher in the presence of MT hints than with the source sentence only, indicating that the translation system successfully aids gist comprehension. Somewhat surprisingly, participants performed better given only the MT hint than when they had both the MT and source sentences for assistance. This mirrors the results of O'Regan et al. (2013) who proposed that it was due to an 'information glut effect', in which participants are overwhelmed or confused when presented with both hints.

| Hint combination | Correct % |
|---|---|
| Reference only | 33.09 |
| Reference + source | 66.13 |
| **Reference + system** | **80.92** |
| Ref. + source + sys. | 77.38 |

Table 3: Mean percentage of gaps correctly filled for the four hint conditions.

A Repeated Measures ANOVA analysis of participants' scores yields $F(3, 21) = 48.229,\ p < 0.01$, indicating that

docs.google.com/forms/d/1-hx2a0pXmIoT2M0sI5okbxkY3hd0zMbFPz0QhSiQ3f4/viewform and docs.google.com/forms/d/1AH7fgdhkGkV6J6vkY6bQh4CuzQnWQtn5FJUoVmtUOKA/viewform

[15] http://wiki.apertium.org/wiki/Translation_quality_statistics

these results are statistically significant.
At 80.92% correctly filled gaps the `eng-sco` MT system compares extremely favourably with results reported for other language pairs Basque-English (`eu-en`), Basque-Spanish (`eu-es`), and Tatar-Russian (`tat-rus`) (Fig. 4), though those are of course cross-language family translations.[16]

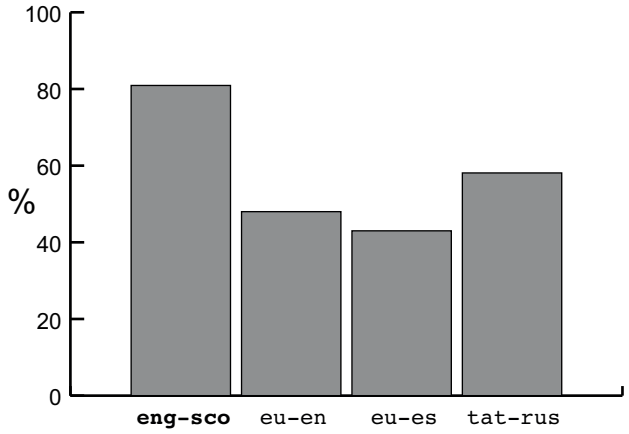


Figure 4: Comparison of assimilation evaluation results (Reference + system gapfill) for `eng-sco` and those reported for other `Apertium` language pairs.

## 6. Error analysis

Comparing source, reference and MT system translation sentences, some failures of the current system are apparent. At under 3000 Scots lexical items the system still lacks a lot of vocabulary and many words are simply not found in the `sco-sco` dictionary, and therefore not translated e.g., *wonderin* ('wondering'), *oer* ('over'), *birl* ('whirl'), *raip* ('rope').
Idiomatic phrases have, on the whole, not yet been incorporated into the system, so constructions such as *fashed hirsell* or *gein it laldy* are treated as individual words rather than translated correctly as 'worried' and 'going at it' respectively.
In many cases, while an item may be present in the dictionaries, spelling variation in the input leads to non-translation e.g., *shakk* instead of *shak* ('shake'), *tel* rather than *tell* ('tell'), *awfy* in place of *awfi* ('awful').
Where two or more options are possible, the system sometimes selects the wrong translation. In some cases this is due to error made by the POS tagger e.g., *tae* is mistranslated as *too* rather than *to*. In others, when presented with options within a POS category, the system fails to select the most appropriate one given the context e.g., *scud* is translated as *smack*, whereas *drink* would be more appropriate given the context.

## 7. Conclusion & future work

As a result of this project a first MT system has been developed for Scots and English, and initial results compare favourably with those published for other language pairs within the platform.
There is however much scope for improvement of the system. Future development will focus on expanding its lexical range and enabling it to better deal with spelling variation by greatly increasing the size of the dictionaries. Another step will be to develop a corpus with which to retrain the POS tagger, which should result in more accurate translations. Further improvements could be made in the disambiguation of words with multiple translations by adding rules to the constraint grammar module.
Subsequent future work will focus on translating from English to Scots, which has not yet been evaluated and would enable learners and non-Scots speakers to produce text and engage with the language, and a greater range of Scots texts to be generated.

[16] Indeed, even the *Ref. + src* score (with no MT hint) for `eng-sco` beats the *Ref. + sys* figures of these language pairs, highlighting the difference in difficulty of the task.

## 8. Acknowledgements

I would like to thank Jacob Nordfalk, Jürgen Wedekind, and all the volunteer evaluation participants for their help with this project.

## 9. References

Ageeva, E., Tyers, F. M., Forcada, M. L., and Pérez-Ortiz, J. A. (2015). Evaluating machine translation for assimilation via a gap-filling task.

Beal, J. (1997). Syntax and morphology. *The Edinburgh history of the Scots language*, pages 335–77.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Forcada, M. L. (2006). Open source machine translation: an opportunity for minor languages. In *Proceedings of the Workshop 'Strategies for developing machine translation for minority languages', LREC*, volume 6, pages 1–6. Citeseer.

Gonzàlez, M., Giménez, J., and Màrquez, L. (2012). A graphical interface for MT evaluation and error analysis. In *Proceedings of the ACL 2012 System Demonstrations*, pages 139–144. Association for Computational Linguistics.

Hutchins, J. (2003a). The development and use of machine translation systems and computer-based translation tools. *International journal of translation*, 15(1):5–26.

Hutchins, J. (2003b). Machine translation and computer-based translation tools: What's available and how it's used. *A New Spectrum of Translation Studies. University of Valladolid*.

Jones, C. (1997). *The Edinburgh history of the Scots language*. Edinburgh University Press.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 605. Association for Computational Linguistics.

Macafee, C. and Aitken, A. (2002). A history of Scots to 1700. *A dictionary of the Older Scottish Tongue*, 12.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

O'Regan, J., Forcada Zubizarreta, M. L., et al. (2013). Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on apertium.org.

Sánchez-Martínez, F. and Forcada, M. L. (2009). Inferring shallow-transfer machine translation rules from small parallel corpora. *Journal of Artificial Intelligence Research*, pages 605–635.

Trosterud, T. and Unhammer, K. B. (2012). Evaluating North Sámi to Norwegian assimilation RBMT. *Free/Open-Source Rule-Based Machine Translation*, 14:13.

Tulloch, G. (1997). Lexis. *The Edinburgh history of the Scots language*, pages 378–432.