Rule-based Breton to French machine translation

Francis M. Tyers
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03070 Alacant
ftyers@dlsi.ua.es

Abstract

This paper describes a rule-based machine translation system from Breton to French intended for producing *gisting* translations. The paper presents a summary of the ongoing development of the system, along with an evaluation of two versions, and some reflection on the use of MT systems for lesser-resourced or minority languages.

1 Introduction

This paper describes the development of a "gisting" machine translation system between Breton and French.[1] The first section will give a general overview of the two languages in question and describe the aims for the current system. The subsequent sections will describe some of the development work, the current status, an evaluation, and some prospects for future development.

Breton is a Celtic language of the Brythonic branch which is today largely spoken in Brittany in the north-west of France. Historically it was spoken to different degrees throughout Brittany, but has been losing territory to French since the 12th century, most rapidly in the last 100 years. The language is classed as a language in "serious danger of extinction" by the *UNESCO Red Book on Endangered Languages* (Salminen, 1999).

Although both Breton and French both belong to the large Indo-European language group, they are from widely different families, Celtic and Romance respectively, with a number of substantial differences between them.

[1]The system can be tried out online at http://www.ofis-bzh.org/bzh/ressources_linguistiques/index-troerofis.php

Speakers of minority or regional languages typically differ from the majority in being bilingual, speaking both their language, and the language of the majority. In contrast, a majority language speaker does not usually speak the minority or regional language. This has some implications for the requirements society will put on machine translation systems.

Applications of machine translation system can be divided in two main groups: *assimilation*, that is, to enable a user to understand what the text is about; and *dissemination*, that is, to help in the task of translating a text to be published. The requirements of either group of applications is different.

Assimilation may be possible even when the text is far from being grammatically correct; however, for dissemination, the effort needed to correct (*post-edit*) the text must not be so high that it is preferable to translate it manually from scratch.

A majority to minority language system will mainly be used for dissemination purposes; it must therefore be such that post-editing the output is faster than translating from scratch. Intelligibility is secondary, and only important if it helps the post-editor.

A minority to majority language system will, however be mainly used for assimilation, for instance, to answer vital questions such as "what are they writing about me in the minority language newspaper?". Therefore, the main goal is intelligibility.

The system in this paper was indeed developed with this second objective in mind, to be able to provide intelligible translations into French of text published in the Breton language media, and to attempt to decrease the need to translate everything published in Breton into French in order to be understood by interested people who do not speak

Breton.

Two examples of this are: A shared e-mail conversation between Breton speakers, with one French speaker joining in. As opposed to having the whole conversation translated, it should be possible to use the MT system to get a *gist* and then request clarification. Another would be to allow Breton-speaking organisations to keep notes of meetings and minutes in Breton, while not excluding members who do not speak Breton from understanding what is going on.

There have been two releases of the system (`apertium-br-fr`), the first one is version 0.1, which was released in May of 2009, and the second is version 0.2, released in March of 2010. Also mentioned in the paper are the results from the prototype *word-for-word* system that was presented in Tyers (2009).

2 Development

The system is based on the Apertium machine translation platform.[2] The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other language pairs, and in particular languages from the Celtic group, e.g. Welsh (Tyers and Donnelly, 2009). The whole platform, both programs and data, are licensed under the Free Software Foundation's General Public Licence[3] (GPL) and all the software and data for the 22 supported language pairs (and the other pairs being worked on) is available for download from the project website.

The initial development time (for version 0.1 of the system) was approximately six person months from scratch. Three of these taken up with development of the morphological analyser, two in development of the bilingual dictionary and one month for the transfer rules. The majority of the work was done by a PhD student of machine translation, with input from Breton speakers. One during the development of the dictionaries, and two during development of the transfer rules. Subsequent expansion of the system has been continued by one PhD student working on the transfer rules, and one Breton speaker working on the dictionaries.

[2] `http://www.apertium.org`

[3] `http://www.fsf.org/licensing/licenses/gpl.html`

2.1 Morphological analyser

As no free morphological analyser for Breton was previously available, a new analyser was written for the project. Words were added to the analyser based on frequency, with the most frequent words being added first. The frequency list was taken from a database dump of the Breton Wikipedia. Some open categories, for example nouns and adjectives were semi-automatically extracted from the French and Breton Wiktionaries.[4]

For example, for nouns, the French Wiktionary has 1,704 entries and the Breton Wiktionary has 2,415 entries. For adjectives the French Wiktionary has 434 entries while the Breton Wiktionary has 557 entries. There are however overlaps and duplicates. No record was kept of how many of these entries finally made it into the analyser as the data was merged with other free data, for example from the Breton–French FreeLang dictionary.[5]

As described below, verb conjugation is very regular, and thus, the main verb paradigms were entered by hand, and entries were largely able to be assigned to paradigms based only on their infinitive. Some verbs feature stem internal variation, and these were added manually.

Closed categories, including pronouns, auxiliary verbs, prepositions, conjunctions etc. were added from two descriptive grammars of Breton (Hemon, 2007; Press, 1986) by hand. The total development time for the morphological analyser for version 0.1 was approximately three person months. The final tagset includes 63 unique tags (for part-of-speech and other morphological properties such as person, number, gender and tense) and 202 unique combinations of tags.[6]

Breton is a weakly inflected language, with the following characteristics. Nouns have two grammatical genders, masculine and feminine and two numbers, singular and plural. The gender of the noun plays an important part in the system of initial consonant mutations. For example, feminine singular nouns mutate after the definite article, *bro* 'country' and *ar vro* 'the country', but masculine singular nouns do not, *mor* 'boat' *ar mor* 'the sea'. Adjectives inflect for comparative, superlative and

[4] `http://fr.wiktionary.org` and `http://br.wiktionary.org/`

[5] `http://meskach.free.fr/arbo/dico/tomaz.html`

[6] A description of the tagset can be found at `http://wiki.apertium.org/wiki/Breton#Tagset`.

exclamative.

Verbs inflect for four persons (first, second, third and impersonal) and two numbers (singular and plural), although the impersonal form of the verb does not have number. There are four main finite tenses: present, imperfect, past definite, and future, and four moods: indicative, conditional, habitual and imperative. The habitual only appears in two verbs, *bezañ* 'to be' and *kaout* 'to have'. The verb *bezañ* in addition has special locative forms, compare *Pelec'h emañ?* 'Where is he?' with *Piv eo?* 'Who is he?'. When the subject is placed before the verb, then the verb is inflected in the third person singular, regardless of the person and number of the subject. There are also a number of non-finite forms of the verb, the infinitive and the past participle.

Verbal inflection is very regular, unlike in French or Spanish, there is only one conjugation class for all verbs. There are however distinct forms of the infinitive, *-iñ*, *-añ*, *-out*, etc. which vary substantially on dialect or region. In the dictionaries, 54 inflectional paradigms which cover the vast majority of verbal inflection patterns, taking into account stem-internal variation, along with different infinitival forms. This can be compared with the 107 used in the French analyser and 118 in the Catalan one). It is worth noting that in addition to the inflected verb forms, Breton also has many periphrastic tense forms, with auxilliaries (*bezañ* 'to be', *ober* 'to do' and *kaout* 'to have').

The cardinal numbers 1–4 inflect for gender (e.g. *daou* 'two+MASC', *div* 'two+FEM') and agree in gender with the noun phrase they modify. Unlike in French, following a plural number, the noun is given in the singular, not plural. Prepositions do not inflect, although they do contract substantially with object pronouns. For example, *ouzh* 'against' can be found in a number of forms with an attached pronoun, *ouzhin* 'against me', *outañ* 'against him', *ouzhimp* 'against us', etc.

The other parts of speech, adverbs, conjunctions, verbal particles[7] etc. do not inflect. A comprehensive description of Breton grammar, in English may be found in Press (1986).

Some basic statistics for the analyser can be found in table 1, and the coverage over two corpora can be found in section 3, evaluation. The large increase in surface forms between the two versions, for a relatively small number of new lemmata can be explained by the number of new multiword verbs (930 in version 0.2 compared to 204 in version 0.1).

Version	Lemmata	Sur. forms	Ambig.[8]
2009-01-01	9,961	105,697	1.07
0.1	11,559	224,839	1.09
0.2	14,693	454,479	1.11

Table 1: Basic statistics of two versions of the Breton morphological analyser. The two versioned numbers are quality checked releases, the 2009-01-01 version is the analyser as used in Tyers (2009).

2.2 Part-of-speech tagger

The part-of-speech tagger for the system is based on two technologies, the first is constraint grammar (Karlsson et al., 1995), which uses linguist-written rules to disambiguate morphologically ambiguous words based on sentence context. The second is a bigram HMM part-of-speech tagger included in the Apertium distribution. The HMM tagger was trained in an unsupervised manner on a database dump of the Breton Wikipedia.[9] The accuracy of the POS tagging has not been evaluated. The constraint grammar is run before the HMM-based part-of-speech tagger and tries to decrease or remove ambiguity where accurate rules can be made. Some of the rules are written in an *ad hoc* manner, treating specific disambiguation problems with specific words. There are also a number of rules which treat lexical disambiguation problems, for example the word *pediñ* can be translated transitively as *inviter* 'invite' or intransitively as *prier* 'pray'.

Version 0.1 of the constraint grammar had 134 rules, while version 0.2 has 206. Two example rules, one treating disambiguation and the other treating lexical selection can be found in figure 1.

2.3 Transfer lexicon

The open categories (nouns, verbs, adjectives, adverbs) in the transfer lexicon, or bilingual dictionary are primarily based on the FreeLang Breton–

[7] Words that come before the verb to indicate subject/object position, mood, tense, etc.

[8] Ambiguity is the mean number of analyses returned for each surface form.

[9] `http://br.wikipedia.org`; Accessed: 10-04-2009

```
LIST BOS    = (>>>) (sent) ;

LIST DetPos = (det pos) ;

LIST Vbloc  = (vbloc) ;
LIST Vblex  = (vblex) ;
LIST Vbser  = (vbser) ;

SET Verb    = Vbloc | Vblex | Vbser ;

# "Ma vez klasket sevel abadennoù"
# Remove a possessive at the beginning
#   of a sentence if it is only followed
#   by a verb.
REMOVE DetPos IF (-1 BOS) (1C Verb) ;

# "ma zud"
# Choose 'parents' as a translation of
#   'tud' instead of 'people' if it is
#   preceeded only by a possessive.
SUBSTITUTE (n) (n :1) ("den"ri n m pl)
                      (-1C DetPos) ;
```

Figure 1: Two Constraint Grammar rules for Breton. The first is for morphological disambiguation, and the second for lexical selection. The comments (lines preceeded with '#'), give an example sentence first in quotes, followed by a description of the rule.

French bilingual wordlist.[10] From this, entries were extracted which were covered by both the Breton analyser and the French analyser, and then where they were unambiguous (e.g. noun or adjective on both sides) they were added to the lexicon. The closed categories (prepositions, pronouns, conjunctions, determiners etc.) were added by hand. The severe lack of parallel corpora for Breton (see e.g. Tyers (2009)) made it very difficult to use any automatic method for extracting vocabulary. Entries in the transfer lexicon take the form of 1:1 pairs, but may include multiword units. Where a word may have more than one translation, either the most frequent or the most general translation is taken. This is motivated by the fact that the platform as of yet does not have any standard method for selecting between ambiguous entries. This will be revisited in section 4.

Version 0.1 of the translator had a transfer lexicon with 11,751 entries, and version 0.2 has 14,549.

2.4 Transfer rules

Structural transfer rules are specified in XML. The structural transfer process is split into three parts. These are:

- The first stage (**chunker**) performs lexical transfer and local syntactic operations and segments the sequence of lexical units into *chunks*. A chunk is defined as a fixed-length sequence of part-of-speech tags that corresponds to some syntactic feature, for example a chunk might encompass all or part of a noun phrase.

- The second stage (**interchunk**) performs more global operations on and between chunks.

- The third stage (**postchunk**) performs another round of local operations inside each chunk and outputs the word stream in the format accepted by the morphological generator.

Table 2 presents details of the number of rules in each version of the system. Space does not permit a full description of all transfer rules, but an overview will be given below.

2.4.1 Conjunctive genitive

The conjunctive genitive in Breton represents probably one of the biggest structural differences between Breton and French. In French, the genitive is formed using the preposition *de*, *La fille du docteur* 'The daughter of the doctor'. In Breton the equivalent construction is to place the two nouns next to each other with the definite article *ar* 'the', as in (1), with neither noun marking for case, e.g. *Merc'h an doktor*. In the examples, the first line gives the Breton, the second line the gloss in English, and the last line the French translation, the English translation is footnoted, this results in a large number of footnotes, but is included in the hope that it helps those who do not understand French.[11]

(1) a. Dor ar skol
Door the school
'La porte de l'école'[12]

b. Dor sal ar skol
Door room the school
'La porte de la salle de l'école'[13]

[10] http://www.freelang.com/dictionnaire/breton.html; Accessed: 10-04-2009

[11] The following abbreviations are used: PART 'Particle', SG 'Singular', PL 'Plural', P1 'First person', P3 'Third person', NEG 'Negative particle', INF 'Infinitive', GER 'Gerund particle', PP 'Past participle'.

[12] 'The door of the school', 'The school's door'

[13] 'The door of the room of the school', 'The school room's door'

Approximately one third of the first-stage transfer rules deal with these and related phenomena, e.g. moving adjective location (some French adjectives are placed before the head and others after, in Breton nearly all are placed after, with the exception of superlative forms which can be placed before or after) and changing demonstratives (Breton demonstratives, are placed at the end of the noun phrase, which starts with a definite article, in French, the demonstrative replaces the definite article and is placed at the beginning of the phrase).

2.4.2 Pronoun placement and insertion

In Breton, pronouns for indirect and direct objects are placed after the verb,[14] always in the same order (verb, direct object, indirect object). In French, the position of the clitic pronouns changes depending on the person and number of the pronouns and the tense of the verb. This can be seen in example (2).

(2) a. Kinnig ac'hanout dezhañ
Present+INF you to him
'Te présenter à lui'[15]

b. Kinnig anezhañ dit
Present+INF him to you
'Te le présenter'[16]

Breton is also a *pro-drop* language, meaning that the pronominal subject of the verb can be dropped. Transfer rules were written to introduce the pronominal subject depending on the person and number of the verb, where no previous subject was seen.

(3) a. Ne implij ket
NEG employ+P3.SG NEG
'Il n'emploie pas'[17]

b. Debriñ a ran
Eat PART do+P1.SG
'Je mange'[18]

2.4.3 Verb tenses

Although both Breton and French have a similar complement of verbal tenses and moods, these are used differently in each language. Transfer rules were written to deal with converting, e.g. periphrastic tenses to inflected tenses, and inflected tenses to different inflected tenses. Example (4-a) shows a gerund in Breton being paraphrased as *en train de* in French. In (4-b) we see a change in auxiliary verb in French, from *être* 'to be' to *avoir* 'to have'.

(4) a. O kanañ kreñv emañ
GER sing+INF loud is
'Il est en train de chanter fort'[19]

b. Lazhet eo bet
Kill+PP is be+PP
'Il a été tué'[20]

c. Ret eo dit dont
Necessary is to you come+INF
'Il faut que tu viennes'[21]

Example (4-c) transfers a Breton structure indicating obligation, formed from a noun, the verb 'to be', an indirect object pronoun and infinitive to a French structure with *Il faut que* and a subjunctive. This rule is split into two stages, the *chunker* takes the *Ret eo* 'Necessary is' and an indirect object pronoun and outputs *Il faut que* and the corresponding subject pronoun. The infinitive is passed on as is. There is then an *interchunk* rule which takes *Il faut que* followed by a subject pronoun and an infinitive and replaces the infinitive with a present subjunctive which agrees with the subject pronoun.

2.4.4 Constituent re-ordering

Breton has fairly free word order, the order of sentences can be VSO (5-b), SVO (5-c) or OVS (5-a), where French is uniformly SVO.[22]

(5) a. Bras eo ar paotr
Big is the boy
'Le garçon est grand'[23]

b. Emañ ar bugel o tebriñ
Is the child GER eat+INF
'L'enfant est en train de manger'[24]

[14] Excepting more formal or dialectal language, where the direct object pronoun may be placed before the verb.
[15] 'To present you to him'
[16] 'To present him to you'
[17] 'He doesn't employ'
[18] 'I eat'
[19] 'He is singing loudly'
[20] 'He has been killed'
[21] 'You have to come', 'You must come'
[22] In these abbreviations S stands for Subject, V stands for Verb and O stands for Object.
[23] 'The boy is big'
[24] 'The child is eating'

Version	Stage	Rules	Total
0.1	I CHUNKER	133	198
	II INTERCHUNK	63	
	III POSTCHUNK	2	
0.2	I CHUNKER	153	222
	II INTERCHUNK	67	
	III POSTCHUNK	2	

Table 2: Number of rules in each of the distinct transfer stages for both released versions of the translator.

c. Ni a ro al laezh da Vari.
We PART give the milk to Mari.
'Nous donnons le lait à Mari.'[25]

The examples in (5) are all treated correctly in the current version. There are however a large number of cases which are not covered by transfer rules. Refer to section 3.3 for a short discussion on these.

3 Evaluation

There have been two evaluations of the system, corresponding to the two currently released versions, 0.1 and 0.2. The evaluation of the same test set for the *word-for-word* system described in Tyers (2009) is also given for comparison. The systems were evaluated in two ways. The first was the coverage[26] of the system. The second was the word error rate (WER) of the translations produced when comparing with a corrected sentence.

3.1 Coverage

Coverage of the system was calculated over two freely-available corpora of Breton. The first is a database dump of the Breton Wikipedia (WP),[27] and the second is the archive of the online weekly news magazine *Bremaik* (BM).[28]

3.2 Translation quality

The translation quality was measured using two metrics, the first was word error rate (WER), and the second was position-independent word error rate (PER). Both metrics are based on the Levenshtein distance (Levenshtein, 1965) and

Corpus	Tokens	Version	Coverage
WP	2,724,465	0.1	84%
	2,691,517	0.2	87%
BM	397,641	0.1	85%
	390,630	0.2	90%

Table 3: Naïve vocabulary coverage over two available corpora of Breton. Differences in numbers of tokens between the same corpus due to different tokenisation, for example new multiword expressions.

Version	WER	PER
2009-01-01	59%	39%
0.1	41%	23%
0.2	38%	22%

Table 4: Word error rate and position-independent word error rate over the 5,804 word test corpus for three versions of the translator.

were calculated for each of the sentences using the `apertium-eval-translator` tool.[29] Metrics based on word error rate were chosen as to be able to compare the system against systems based on similar technology, and to assess the usefulness of the system in a real setting, that is of translating for dissemination.

A corpus of 398 sentences (5,804 words) was extracted from the Bremaik archives. Sentences were extracted pseudo-randomly, while fitting the following conditions: The sentences did not have any unknown words, and the sentences were between 5–30 words long. The reason for only selecting sentences with no unknown words was to test the quality of the transfer and disambiguation rule-sets as opposed to the coverage of the dictionaries. The test corpus, containing both the original sentences, the two machine translation outputs and the post-editions is publically available for download.[30]

The big difference between the WER and PER scores is due to the fact that the system only does local reordering (for example within the noun phrase). Global, constituent reordering is not done, except for very simple phrases as the system does not permit this easily. This restriction is discussed further in section 4.

These numbers can be compared with other

[25] 'We give the milk to Mari'

[26] Here coverage is defined as *naïve coverage*, that is for any given surface form at least one analysis is returned. This may not be complete.

[27] `http://xixona.dlsi.ua.es/~fran/breton/Wikipedia.br.gz`; Accessed 08-02-2009

[28] `http://xixona.dlsi.ua.es/~fran/breton/Bremaik.tar.gz`; Accessed 08-02-2009

[29] `http://sourceforge.net/project/showfiles.php?group_id=143781&package_id=206517`; Version 1.0, 4th October 2006.

[30] `http://xixona.dlsi.ua.es/~fran/breton/br-test-corpus.tar.gz`

scores for translators in the Apertium platform, for example the Welsh–English system described in Tyers and Donnelly (2009) achieves post-edition WER of 53.40% and PER of 27.22% over 5,392 words. The Basque–Spanish system in Ginestí-Rosell et al. (2009) reports WER of 72.41% and PER of 39.86% over 1,312 words, and to compare with a system between more related languages, the Catalan–Occitan system described in Armentano-Oller and Forcada (2006) achieves a WER of 9.6%. This suggests that the performance is comparable with similar systems between less-related languages.

3.3 Qualitative

Along with the quantitative evaluation of post-edition effort, it is also useful to perform a qualitative evaluation, determining where the system can be improved. Every part of the system can be improved, but it is worth highlighting the following issues which are currently found in the system. In the examples that follow, the Breton phrase is presented on the first line, followed by the current translation produced by the system, which is followed by the correct translation on the third line.

While the disambiguation stage performs well, there are currently some issues, as in example (6). There is an ambiguity between *da* meaning *à* 'to' as a preposition and *ton* 'your' as a possessive. As these appear often in the same position (preceeding a noun phrase), they are difficult to disambiguate. Another difficult word to disambiguate is *re*, as can be seen from examples (6), (7), (8), and (9).

(6) Dedennus eo da labour.
*Intéressant est à travail.
Ton travail est intéressant.[31]

(7) Re ruz am eus c'hoant.
*Celui rouges ai envie.
Je veux des rouges.[32]

(8) Re vras eo ar bragoù.
*Celui grands est ce pantalon.
Ce pantalon est trop grand.[33]

(9) Ar re vras eo ar re wellañ.
*Ceux qui grand il est ceux qui le plus mieux.
Les grands sont les meilleurs.[34]

Complex noun phrases involving co-ordination (10) and sub-ordinate clauses (11) are also not treated.

(10) Emañ ar paotr hag ar plac'h o vont.
*Il est le garçon et la fille en train d'aller.
Le garçon et la fille sont en train d'aller.[35]

(11) Emañ ar plac'h a welan o vont.
*Il est la fille je vois en train d'aller.
La fille que je vois est en train d'aller.[36]

Breton uses the verb *bezañ* 'to be' to form the *existential*, e.g. 'There is' in English. In French this is formed with *il y* and a form of the verb *avoir* 'to have'. This cannot currently be disambiguated between these two uses, and as such produce incorrect translations as in (12). Note that this default method of translating *bezañ* produces correct translations where the usage is not existential as in (13).

(12) Bara *a zo* war an daol.
*Pain *est* sur la table.
Il y a du pain sur la table.[37]

(13) Ar brezhoneg *a zo* ur yezh keltiek.
Le breton *est* une langue celtique.[38]

It is also worth noting that nothing has been yet done to treat translating the imperative. This mood is infrequent in most news text (at which the translator is aimed) and insufficiently adequate disambiguation (many forms are homographs) would decrease the accuracy of the translator.

4 Future work

It is intended to continue the development of the system to further improve the quality of the translations. There are a number of areas where it is believed that more work would yield better results, they are:

- Coverage – Although the coverage of the dictionaries is good, with an increase of coverage of perhaps 5–6%, the translations would be much more intelligible.
- Better source-language disambiguation – In some cases errors in the part-of-speech tagging of Breton cause substantial problems in

[31] 'Your work is interesting.'
[32] 'I want some red ones'.
[33] 'These trousers are too big.'
[34] 'The big ones are the best.'
[35] 'The boy and the girl are going.'
[36] 'The girl that I am looking at is going.'
[37] 'There is bread on the table.'
[38] 'Breton is a Celtic language.'

the translation. In other cases the translation remains intelligible, but suffers in fluency. There are a number of avenues open for improving the part-of-speech tagging, including writing better and more accurate constraint grammar rules, or tagging a corpus of Breton and performing supervised training of the HMM-based tagger in place of unsupervised training, which has been shown to provide more accurate disambiguation.

- Lexical selection – As Breton and French are less closely related than many of the languages which have been dealt with in Apertium up to this point, there are many instances where the selection of a translation results in an inadequate or non-fluent translation. A comprehensive lexical selection module which allows for the bilingual dictionary to have more than one correspondence in the target language for each source language lexical unit could improve the system in this respect.

- Deeper transfer – Because of the way in which the transfer system only works on fixed-length sequences of lexical forms, it is problematic to do long distance re-ordering, for example of relative or subordinate clauses. A deeper approach to transfer, working on parse trees could get around this problem by allowing rules to apply recursively.

A more in depth human evaluation is also planned in order to test the system in the assimilation setting for which it was designed. Along with this testing the system in more real-world conditions (for example, texts containing unknown words) is planned to evaluate how robust the transfer rules are.

5 Concluding remarks

This paper has presented the first rule-based Breton to French machine translation system. It has presented two evaluations showing the performance of the system for the post-edition task improving over time. Initial results are promising, although the system is not yet suitable for producing text for dissemination.

Acknowledgements

Work on version 0.1 of the system was financially supported by Grup Transducens at the Universitat d'Alacant, Ofis ar Brezhoneg, and Prompsit Language Engineering. I am very grateful to *Ofis ar Brezhoneg* for making available their translation memory, and for their consistent help during the project. I would also like to extend special thanks to: Fulup Jakez, the director, for his work on verifying and expanding the Breton morphological analyser and Breton–French lexicon, and to both Fulup and Gwenvael Jekel for providing descriptions of transfer rules. This work has also received the support of the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

References

Armentano-Oller, Carme and Mikel L. Forcada. 2006. Open-source machine translation between small languages: Catalan and Aranese Occitan. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, pages 51–54. organised in conjunction with LREC 2006 (22-28.05.2006).

Ginestí-Rosell, Mireia, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Francis M. Tyers, and Mikel L. Forcada. 2009. Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, (43):187–195.

Hemon, Roparz. 2007. *Breton Grammar*. Evertype. translated by Michael Everson.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Levenshtein, Vladimir I. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. English translation in Soviet Physics Doklady, 10(8), 707–710, 1966.

Press, Ian. 1986. *A Grammar of Modern Breton*. Mouton de Gruyter.

Salminen, Tapani. 1999. *UNESCO Red Book on Endangered Languages*. UNESCO. `http://www.tooyoo.l.u-tokyo.ac.jp/archive/RedBook/index.html`.

Tyers, Francis M. and Kevin Donnelly. 2009. apertium-cy: A collaboratively-developed free RBMT system for Welsh to English. *Prague Bulletin of Mathematical Linguistics*, (91):57–66.

Tyers, Francis M. 2009. Rule-based augmentation of training data for breton–french statistical machine translation. *Proceedings of the 13th Conference of the European Association for Machine Translation*, pages 213–218.