

Rule-based augmentation of training data for Breton–French statistical machine translation

Francis M. Tyers
Universitat d’Alacant

Departament de Llenguatges i Sistemes Informàtics
Universitat d’Alacant
E-03071 Alacant
Spain
email: ftyers@prompsit.com

1 Introduction

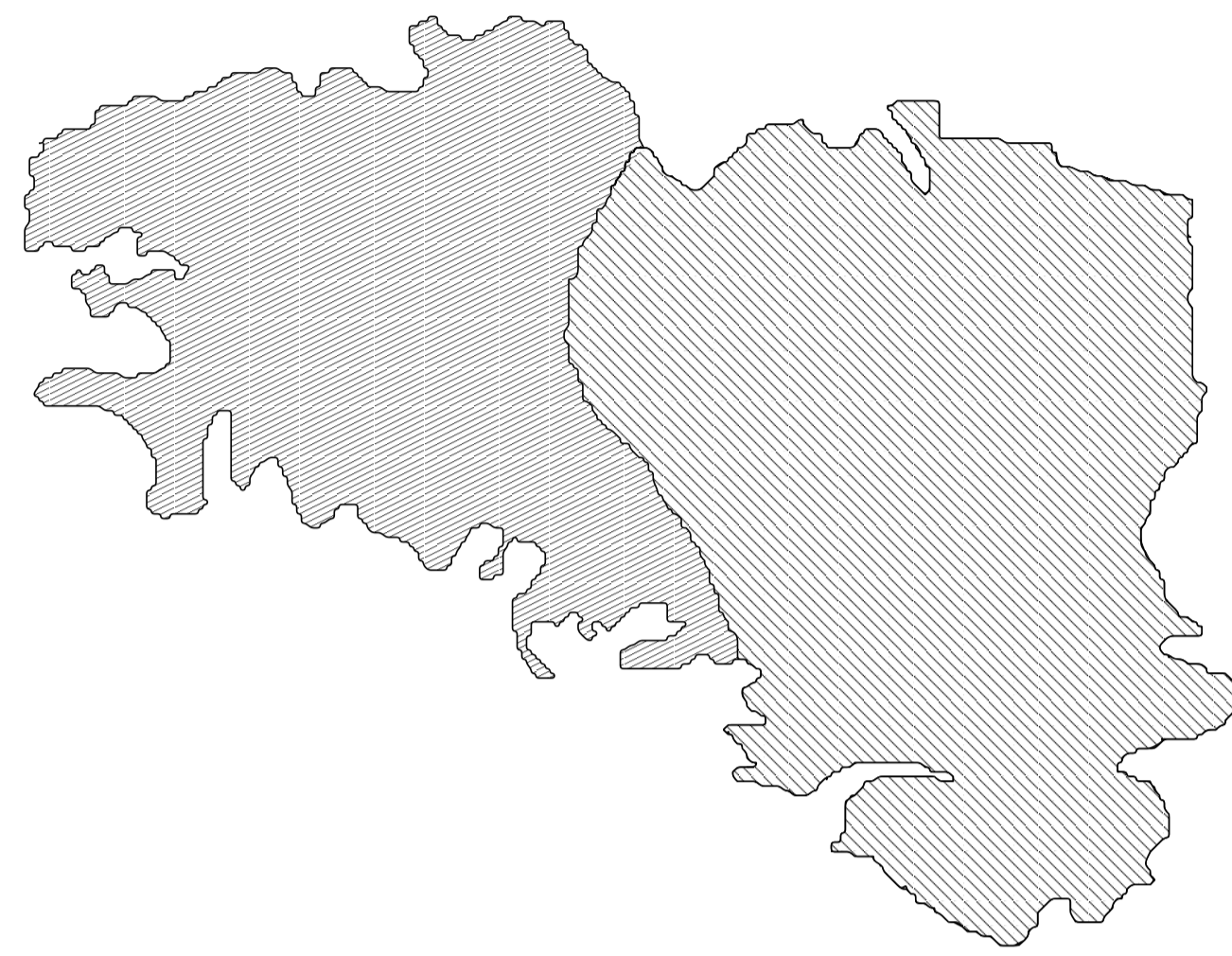


Figure 1 Map of Breizh (Brittany), with the traditionally Breton-speaking area, Breizh-Izel (Lower Brittany) in the west.

Breton is a Celtic language of the Brythonic branch largely spoken in Brittany in the north-west of France. Historically it was spoken only in the northern part of Brittany, **Breizh-Izel** (Lower Brittany). In **Breizh-Uhel** (Higher Brittany), traditionally Romance languages are spoken.

Breton	Welsh	Irish	French	English
ti	tŷ	teach	maison	house
dour	dŵr	uisce	eau	water
mab	mab	mac	fils	son
penn	pen	ceann	tête	head
aval	afal	ubhal	pomme	apple
amzer	amsr	aimsir	temps	time
skrivañ	ysgrifennu	scríobh	écrire	write

Table 1 Comparative table of three Celtic languages, English and French

According to the organisation **Ya d’ar brezhoneg**, the number of native speakers of Breton is around 200,000 as of early 2009. The number is decreasing at a rate of at least one per hour, although the number of new speakers, as a result of bilingual education for children, in the form of **Diwan**, **Div Yezh** and **Dihun** schools and adult education, has been increasing steadily since the 1970s.

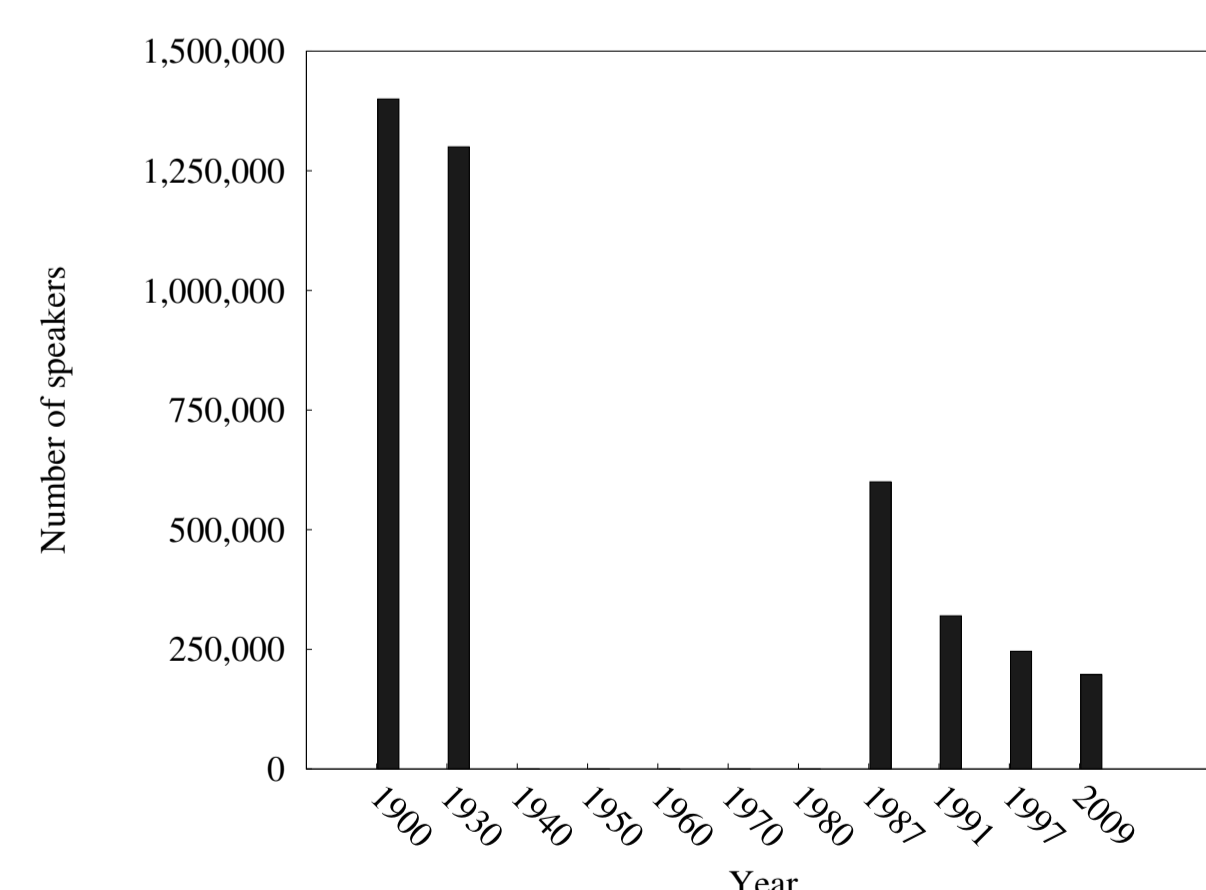


Figure 2 The decline of the Breton speaking population 1900–2009

Like other Celtic languages, Breton exhibits the phenomenon of initial consonant mutation. This occurs when the initial consonant of a word changes based on morpho-syntactic context.

‘father’ → *tad*
‘my father’ → *ma zad*
‘your father’ → *da dad*

Figure 3 The word ‘*tad*’ (father), an example of initial consonant mutation

2 Resources

As for many less-resourced language pairs, there is very little aligned bilingual text.

- Parallel corpora for Breton–*anything* are few and far between.
- Small parallel corpus from the Ofis ar Brezhoneg
 - Breton and French
 - Short segments from a translation memory
 - Largely in the domains of tourism, localisation and education
- 31,000 lines
- Approximately 285,000 Breton words and 282,073 French words

```
<TrU>
<Seg L=FR-FR>à la pointe dans les nouvelles
technologies de l’information et de la
communication, la Bretagne investit pour
inventer le campus numérique
<Seg L=EN-CA>Breizh zo er penn a-raok war dachenn
teknologiezhioù nevez ar c’helaouñ hag ar
c’hehentiñ, ha stagañ a ra ganti da ijin añ
ar c’hampus niverel.
</TrU>
<TrU>
<Seg L=FR-FR> Le campus numérique breton ?
<Seg L=EN-CA> Ar c’hampus niverel breizhat ?
</TrU>
<TrU>
<Seg L=FR-FR>La Région soutient les entreprises
bretonnes
<Seg L=EN-CA>Embregerezhioù Breizh skoazellet
gant ar Rannvro
</TrU>
<TrU>
<Seg L=FR-FR>Dans le contexte de crise actuelle, les
actions régionales de soutien au développement
économique prennent une dimension particulière.
<Seg L=EN-CA>Gant an enkadenn zo bremañ e teu oberoù
ar rannvro evit harpañ diorren an ekonomiezh da
vezañ pouezus.
</TrU>
<TrU>
<Seg L=FR-FR>Naissance de la Bretagne
<Seg L=EN-CA>Ganedigezh Breizh
</TrU>
```

Figure 4 Extract from the Ofis ar Brezhoneg translation memory

3 Two machine translation systems

A rule-based MT system for Breton–French is currently being developed inside the Apertium project (<http://www.apertium.org/>) The current status of the Breton–French system is as follows: the system has a morphological analyser for Breton with approximately 11,000 lemmata (approx. 85% coverage), a bilingual dictionary with 10,797 part-of-speech tagged correspondences, and a very small number of transfer rules (e.g. for concordance and re-ordering within noun phrases, verbal conjugation and pronoun insertion). In the morphological analyser, there are two kinds of paradigms (<par>), the first for specifying the initial consonant mutations, the second for listing all the morphological forms of a given word and their analyses. For verbs, one combination of lemma and paradigm can generate between 37 surface forms (for unmutating initial consonants) and 193 (for mutating initial consonants). Examples for the verbs *labourat* ‘to work’ and *kadarnaat* ‘to confirm’, including inflectional paradigm *labour/at_vblex* and mutation paradigm (*initial-k*) can be found below:

```
<e lm="labourat">
<i>labour</i>
<par n="labour/at_vblex"/>
</e>
<e lm="kadarnaat">
<par n="initial-k"/>
<i>adarna</i>
<par n="labour/at_vblex"/>
</e>
```

Figure 5 Example of morphological analyser entries for two verbs

Figure 6 shows two bilingual lexicon entries for verbs. The bilingual lexicon specifies correspondences between lemmata and parts of speech.

```
<e>
<p>
<l>labourat<s n="vblex"/></l>
<r>travailler<s n="vblex"/></r>
</p>
</e>
<e>
<p>
<l>kadarnaat<s n="vblex"/></l>
<r>confirmer<s n="vblex"/></r>
</p>
</e>
```

Figure 6 Bilingual dictionary entries

A phrase-based statistical model was trained using Moses and a 3-gram language model using IRSTLM. The rest of the training process followed the instructions for the baseline system for WMT08.

4 Extending the parallel corpus

To try and alleviate the problem of low coverage of the training data, the resources available in the nascent rule-based system were used. Two approaches were taken.

- Simply adding the bilingual transfer lexicon from the system to the end of the training data. This consisted of 10,797 lemmata.
- Automatically generating appropriate mappings between all of the surface forms of the given lemmata in the dictionaries of this system. This consisted of 116,514 mappings of inflected Breton forms to inflected French forms.

In order to generate the surface-form mappings, an expansion of all possible surface forms was taken, along with analyses in the Breton morphological analyser. This was then passed through the rest of the Apertium pipeline in order to produce all of the translations of surface forms in French. This produces a bilingual inflected wordlist (see figure 7). Entries for verbs are generated, where appropriate (e.g. finite verb tenses) with the corresponding subject pronoun in French, and Breton tenses which are not found in French are converted into French tenses (e.g. past habitual is converted to imperfect).

```
mignon,ami
mignoned,amis
vignon,ami
vignoned,amis
dale,retarde
dale,il retarde
labouren,je travaillais
...
```

Figure 7 ‘Expanded’ dictionary entries

5 Evaluation and error analysis

As expected, the number of unknown words decreases when the bilingual lexicon is added to the training data, and even more when the fully-expanded bilingual lexicon is added. The rise in BLEU is probably also due to side effects such as a better word alignment and a better French context available to the language model scoring.

System	Description	BLEU	Phrases	Unknown words
system 1	word-for-word	0.16	n/a	1,191
system 2	baseline SMT	0.29	800k	623
system 3	+ uninflected dictionary	0.30	807k	562
system 4	+ inflected dictionary	0.36	843k	531

Example 1	<i>Benveg troc’hañ dre silabennoù</i>
ref	outil de césure par syllabe
gloss	hyphenation tool
system 3	outil coupe par silabennoù
system 4	outil de coupure par syllabes
Example 2	<i>E rankit kevreañ ouzh an holl darzhioù roadennoù</i>
ref	Vous devez vous connecter à toutes les sources de données
gloss	You should connect to all of the data sources
system 3	Devez connexion de données . les darzhioù
system 4	Vous devez se connecter tous les darzhioù de données
Example 3	<i>Emirelezhioù Arab Unanet</i>
ref	émirats arabes unis
gloss	United Arab Emirates
system 3	émirats arabes unies
system 4	émirats arabes uniszez

6 Future work

Improving this work, it might also be possible to try to learn probabilities for the rule-based created phrase pairs as in Koehn and Knight (2000). Another option would be to try and create “expanded” phrases based on chunks extracted from a bilingual corpus. For example if you have *war toenn an ti*, “sur le toit de la maison” (on the roof of the house), it would be fairly straightforward to generate all possible morphological combinations, viz. *war toennoù an ti*, “sur les toits de la maison”, *war toennoù an tiez*, “sur les toits des maisons”, and *war toenn an tiez*, “sur le toit des maisons” respectively.

It is also worth noting that at present the Breton–French lexicon in Apertium has only one (generally the most frequent) translation per word. It would be feasible to generate more than one entry per word, and then score these on language models.

7 Conclusion

The method presented here is knowledge-light, requiring only a morphological analyser, bilingual dictionary and some very basic transfer rules (for verb conjugation) and could be applied to other under-resourced language pairs to improve the coverage of a statistical system where little parallel data is available.

Acknowledgements

Many thanks to the Ofis ar Brezhoneg, in particular its director Fulup Jakez. Thanks also to everyone at Prompsit Language Engineering, Mikel L. Forcada, and two contributors who do not wish to be named.