

# Rule-based Breton to French 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

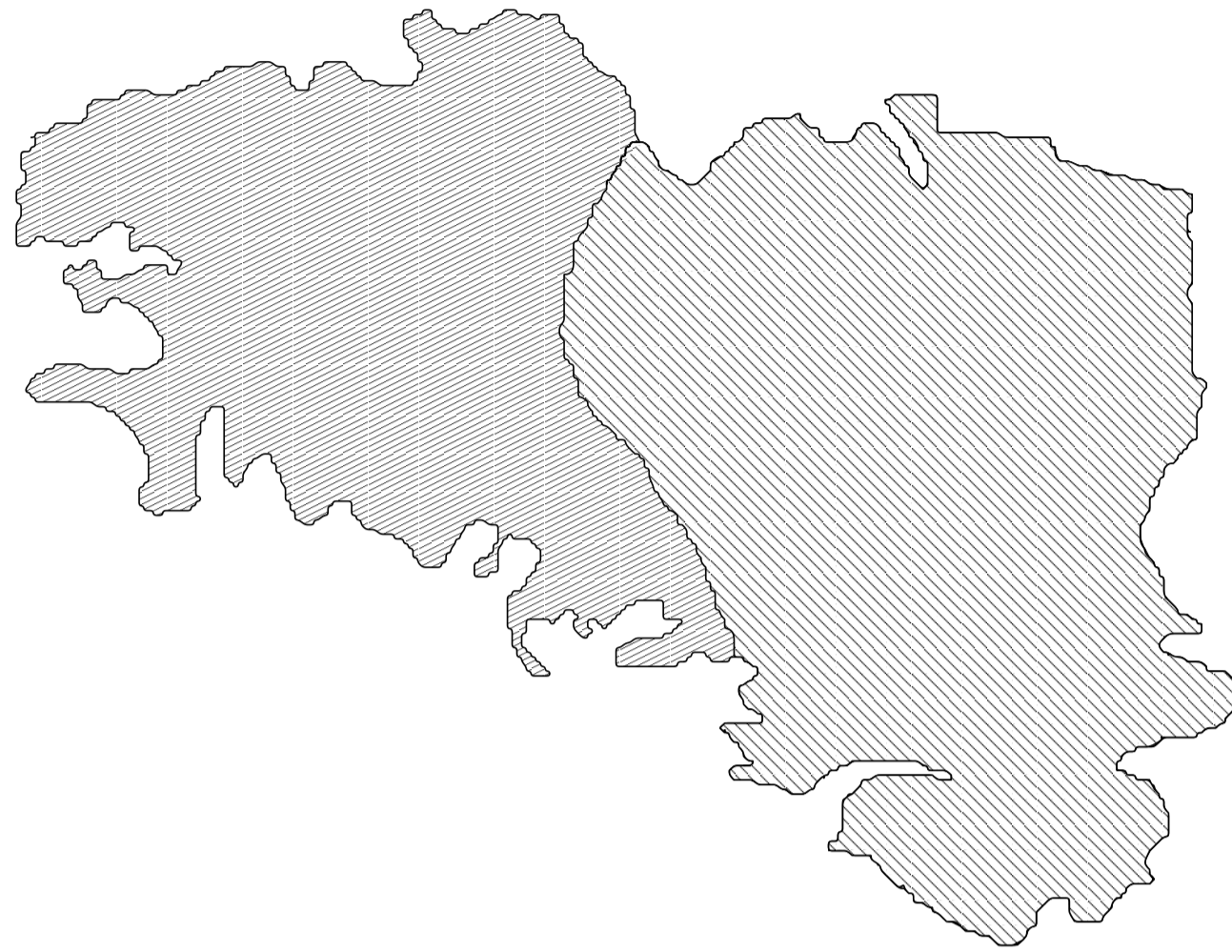


Figure 1: Map of Breizh (Brittany), with the traditionally Breton-speaking area, Breizh-Izel (Lower Brittany) in the west.

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 north-western 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	fil	son
penn	pen	ceann	tête	head
aval	afal	ubhal	pomme	apple
amzer	amser	aimsir	temps	time
skrivañ	ysgrifennu	scríobh	écrire	write

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

According to **Ofis ar Brezhoneg**, the number of native speakers of Breton is around 188,000 as of May 2010. 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.

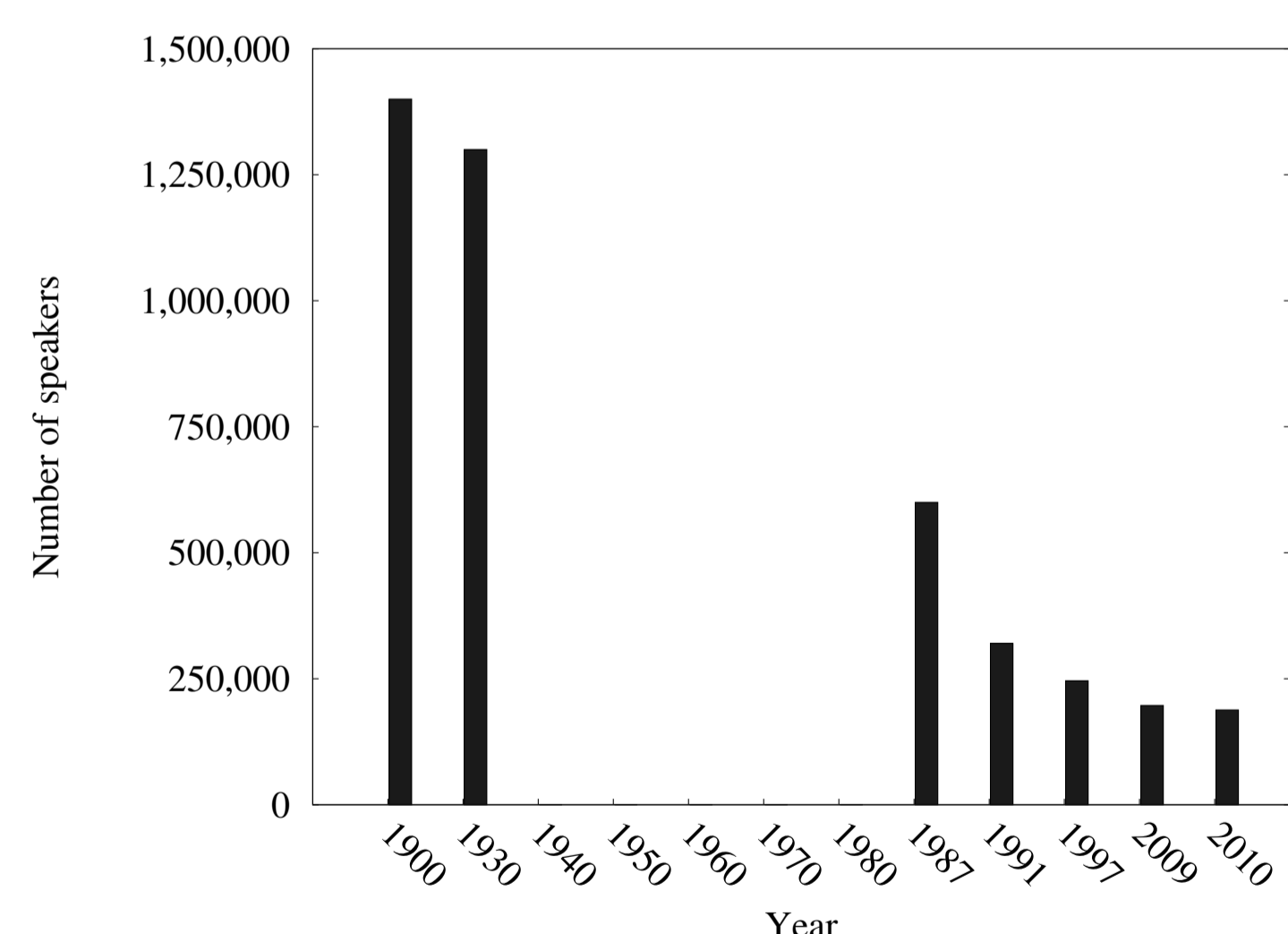


Figure 2: The decline of the Breton speaking population 1900–2010

### 1.1 Why Breton–French machine translation?

To provide broadly intelligible translations of Breton text for French speakers, enabling:

- French speakers to follow the news in the Breton-speaking community
- Communication in Breton without excluding those who do not speak it
  - Organisations to keep minutes and notes in Breton
  - Breton speakers to maintain shared email conversations in Breton
- Vocabulary assistance to learners of Breton

## 2 Components

The system is based on **Apertium** (<http://www.apertium.org/>), a free/open-source rule-based machine translation platform.

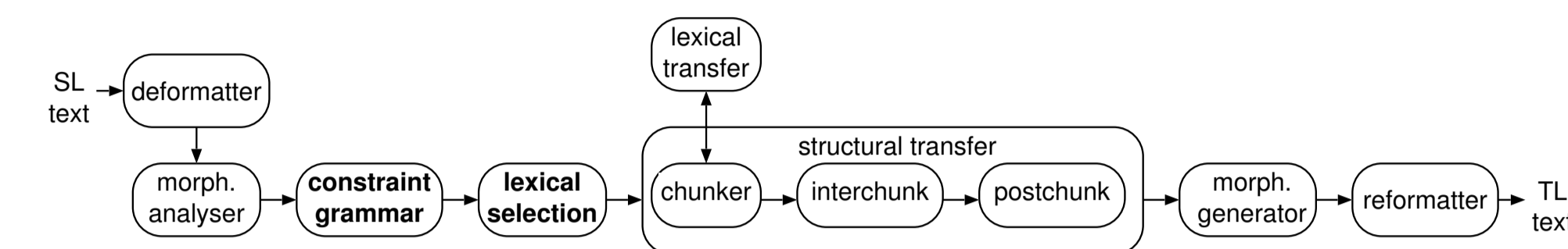


Figure 4: Modules of the Apertium translation system

### 2.1 Morphological analysis

The Breton morphological analyser returns, for every Breton word, the possible lexical forms (analyses) of the word. For the sentence *Gallout a ran ober an dra-se*. ‘I can do that [thing].’

```
^Gallout/Gallout<vblex><inf>$
^a/a<pr>/a<vpart><subj>/mont<vblex><pri><p3><sg>$
^ran/ober<vblex><pri><p1><sg>$
^ober/ober<vblex><inf>/ober<n><m><sg>$
^an/an<det><def><sp>/mont<vblex><pri><p1><sg>$
^dra/tra<n><m><sg>$
^-se/se<adv>$
^./.<sent>$
```

Figure 5: Output of morphological analysis with a finite-state transducer

The analyser has a coverage of between 87–90% over two free corpora of Breton (Wikipedia and *Bremaik*).

### 2.2 Part-of-speech tagging

The part-of-speech tagging module for the system is based on two technologies, the first is Constraint Grammar, which uses linguist-written rules to disambiguate morphologically ambiguous words based on sentence context. The second is a bigram HMM part-of-speech tagger.

#### 2.2.1 Constraint grammar

The Breton constraint grammar has been written manually and contains 206 rules for disambiguating Breton sentences.

```
^Gallout/Gallout<vblex><inf>$
^a/a<vpart><subj>$ SELECT:194
^ran/ober<vblex><pri><p1><sg>$
^ober/ober<vblex><inf>$ REMOVE:205
^an/an<det><def><sp>$ REMOVE:322
^dra/tra<n><m><sg>$
^-se/se<adv>$
^./.<sent>$
```

Figure 6: Morphological disambiguation with constraint grammar

An extract from the constraint grammar rule set:

```
# ex. "Dav e vije"
SELECT:194 Vpart IF (1C VerbFin);

# ex. "Ober a ra gwestell"
REMOVE:205 ("ober" n) IF (NOT -1 Det);

# ex. "An dud a zeu da Roazhon"
REMOVE:322 VerbFin IF (0 Det) (1 NC);
```

#### 2.2.2 HMM-based tagger

The HMM-based tagger was trained on a database dump of the Breton Wikipedia (<http://br.wikipedia.org>) and chooses a single analysis where the constraint grammar does not perform a complete disambiguation.

## 2.3 Bilingual dictionary

The bilingual dictionary, or transfer lexicon contains mappings between lemmas, parts-of-speech and other tags. For example, to indicate to the transfer stage that gender and number need to be inserted, or to indicate a change in a feature.

```
<l>an<s n="det"/></l><r>le<s n="det"/></r><par n="sp_GDND"/>
<l>se<s n="adv"/></l><r>ci<s n="adv"/></r>
<l>tra<s n="n"/><s n="m"/></l><r>chose<s n="n"/><s n="f"/></r>
<l>gallout<s n="vblex"/></l><r>pouvoir<s n="vbmod"/></r>
<l>ober<s n="vblex"/></l><r>faire<s n="vblex"/></r>
```

Figure 7: Extract from the bilingual dictionary

## 2.4 Transfer rules

The structural transfer process is split into three parts. Rules are written in XML (see example in Figure 9).

### 2.4.1 Chunker

Local transfer operations and chunking are performed by the first stage. The output in Figure 8 is generated by two rules. The first takes a sequence of an infinitive verb, followed by a verbal particle and a form of the auxiliary *ober* ‘to do’ and outputs the infinitive verb conjugated according to the auxiliary. The second rule (see Figure 9) takes a sequence of determiner, followed by a noun and a demonstrative adverb and outputs a demonstrative determiner followed by the noun.

```
^Verbcj<SV><vbmod><pri><p1><sg>{^pouvoir<vbmod><pri><4><5>$}$
^faire<SV><vblex><inf><sg>{^faire<vblex><inf>$}$
^det_nom<SN><f><sg>{^ce<det><dem><2><3>$ ^chose<n><2><3>$}$
^punt<sent>{^.<sent>$}$
```

Figure 8: Output from the the first transfer stage

```
<rule comment="REGLA: an dra-se : cette chose">
  <pattern>
    <pattern-item n="det"/>
    <pattern-item n="nom"/>
    <pattern-item n="dem"/>
  </pattern>
  <action>
    <call-macro n="f_concord2">
      <with-param pos="2"/>
      <with-param pos="1"/>
    </call-macro>
    <out>
      <chunk name="det_nom">
        <tags>
          <tag><lit-tag v="SN"/></tag>
          <tag><var n="genero"/></tag>
          <tag><var n="numero"/></tag>
        </tags>
        <lu>
          <lit v="ce"/>
          <lit-tag v="det.dem"/>
          <clip pos="1" side="t1" part="gen" link-to="2"/>
          <clip pos="1" side="t1" part="nbr" link-to="3"/>
        </lu>
        <b pos="1"/>
          <lu>
            <clip pos="2" side="t1" part="lem"/>
            <clip pos="2" side="t1" part="a_nom"/>
            <clip pos="2" side="t1" part="gen" link-to="2"/>
            <clip pos="2" side="t1" part="nbr" link-to="3"/>
          </lu>
        </chunk>
      </out>
    </action>
  </rule>
```

Figure 9: Rule to translate a demonstrative noun phrase

### 2.4.2 Interchunk

The second stage of transfer, *interchunk* performs global operations between chunks. For example the insertion and concordance of a missing subject pronoun as in Figure 9.

```
^Prnperssubj<SN>{^je<prn><tn><p1><mf><sg>$}$
^verbcj<SV><vbmod><pri><p1><sg>{^pouvoir<vbmod><pri><4><5>$}$
^faire<SV><vblex><inf><sg>{^faire<vblex><inf>$}$
^det_nom<SN><f><sg>{^ce<det><dem><2><3>$ ^chose<n><2><3>$}$
^punt<sent>{^.<sent>$}$
```

Figure 10: Output from the second transfer stage

### 2.4.3 Postchunk

The third stage performs cleanup operations and merges linked tags.

```
^Je<prn><tn><p1><mf><sg>$
^pouvoir<vbmod><pri><p1><sg>$
^faire<vblex><inf>$
^ce<det><dem><f><sg>$ ^chose<n><f><sg>$
^.<sent>$
```

Figure 11: Output of the third stage of transfer

## 2.5 Morphological generator

The morphological generator for French takes a sequence of lexical forms and generates the appropriate surface forms.

```
Je peux faire cette chose.
```

Figure 12: Output of the morphological generator

## 3 Evaluation

The evaluation used Word error rate (WER) and position-independent word error rate (PER). A corpus of 398 sentences (5,804 words) was extracted from the *Bremaik* archives. Sentences were extracted fitting the following conditions: No unknown words and between 5–30 words long.

Version	WER	PER
word-for-word	59%	39%
apertium-br-fr 0.1	41%	23%
apertium-br-fr 0.2	38%	22%

Table 2: Word error rate and position-independent word error rate

The big difference in the scores for WER and PER is because only local reordering is performed, constituent reordering is reserved for simple phrases.

## 4 Future work

The performance of the system could be improved by:

- Improving coverage
- Better source language disambiguation
- Lexical selection
- Deeper transfer

## Acknowledgements

Funded by: Grup Transducens, Ofis ar Brezhoneg, and Prompsit Language Engineering. Many thanks to the Ofis ar Brezhoneg, in particular its director Fulup Jakez for his work on the system.