# A prototype dependency treebank for Breton

Francis M. Tyers[1] Vinit Ravishankar[2]
(1) School of Linguistics, Higher School of Economics, Moscow
(2) Institute of Formal and Applied Linguistics, Charles University in Prague, Prague
ftyers@hse.ru, vinit.ravishankar@gmail.com

## Résumé

Cet article décrit le développement du premier corpus syntaxiquement annoté de breton. Le corpus fait partie du projet «Universal Dependencies». Dans cet article, nous décrivons la préparation du corpus, certaines constructions spécifiques au breton qui avaient besoin d'un traitement spécial et nous donnons des résultats de l'analyse syntaxique de breton par un nombre d'analyseurs syntaxiques.[1]

## Abstract

**A dependency treebank for Breton**

This paper describes the development of the first syntactically-annotated corpus of Breton. The corpus is part of the Universal Dependencies project. In the paper we describe how the corpus was prepared, some Breton-specific constructions that required special treatment, and in addition we give results for parsing Breton using a number of off-the-shelf data-driven parsers.

Mots-clés : breton, analyse syntaxique de dependences, banque d'abres syntaxiques.

Keywords: breton, dependency parsing, treebank.

# 1 Introduction

Treebanks, or collections of sentences annotated according to some schema, have existed for decades, under a variety of standards intended to represent various details. The Universal Dependencies project (Nivre *et al.*, 2016) is a multilingual collection of treebanks annotated with dependency relations; at the time of writing of this paper, it consists of 115 treebanks in 65 languages. The project aims to enable a cross-linguistically valid schema of dependency annotation, and heavily depends on public contribution of mostly open resources. The existence of this unified collection of treebanks has led to extremely simplified parser creation and evaluation, as exemplified in the CoNLL 2017 shared task on dependency parsing (Zeman *et al.*, 2017).

This paper describes a treebank for Breton, a language spoken in Brittany in the north-west of France. The treebank will be included in the CoNLL 2018 shared task on dependency parsing[2], and we expect that it would provide a starting point for further annotation of Breton. The treebank is the second Celtic language treebank, and the first treebank for a language of the Brythonic subgroup of Celtic, in

[1] Berrskrid: Deskrivañ a ra ar pennad-mañ savidigezh ar c'horpus kentañ bet notennet e ereadurezh e brezhoneg. Ul lodenn eus ar raktres «Universal Dependencies» eo ar c'horpus-se. En teuliad e teskrivomp penaos e oa bet prientet ar c'horpus ha penaos e oa bet pledet gant frammoù dibar zo eus ar brezhoneg. Ouzhpenn-se, reiñ a reomp disoc'hoù dezrannadur ereadurel ar brezhoneg gant dezrannerioù ereadurel zo.

[2] http://universaldependencies.org/conll18/

Universal Dependencies.

The paper is laid out as follows, in Section 2 we give a brief sociolinguistic and typological overview of the Breton. Then in Section 3 we describe some prior work on computational resources and tools for Breton. In Section 4 we describe the composition of the corpus, and in Section 5 we describe some details of the annotation guidelines, paying attention to Breton-specific phenomena. Section 6.1 reports on a small experiment with three popular data-driven parsers, and is followed by some avenues for future work in Section 7 and conclusions in Section 8.

# 2 Breton

Breton (in Breton *brezhoneg*) 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). For an overview of Breton grammar, see Ternes (2008) and for full grammars see Press (1986) and Hemon (2007).

The language has two grammatical genders (masculine and feminine), two numbers (singular and plural)[3] and like the other Brythonic languages has lost the case system. Like other Celtic languages, Breton has contractions of pronouns and prepositions,[4] for example *ganin* 'with me' and *ganit* 'with you' (from *gant* 'with'). Unlike other Celtic languages, Breton also has an indefinite article *un* 'a', and an analytic passive construction with the verb *bezañ* 'be'.[5] Breton also exhibits the initial consonant mutation typical to Celtic languages, and has a fusional morphological system.

Syntactically, Breton has flexible constituent order within the sentence; VSO, SVO and OVS are frequently used with VSO — the classic Celtic order — being the most prominent. Adjectives follow nouns while other modifiers (adjectives, determiners) precede them. Verbs inflect for person, number, tense and mood. Auxiliaries may follow or precede the main verb.

# 3 Related work

There has been very little work to date on natural language processing for Breton. Among the related articles we may find Tyers (2009) who use a morphological analyser and bilingual dictionary to generate training data for statistical machine translation, and Tyers (2010) who describe a free/open-source rule-based machine translation system for Breton to French. There has been more recent work by Poibeau (2014) on treating initial consonant mutations with finite-state transducers. There are a number of written grammars of Breton, we have particularly relied on Press (1986) and Hemon (2007). Whilst no Breton treebanks exist that we know of, there are two existing treebanks for one Celtic language, viz. Irish (Lynn & Foster, 2016; Lynn *et al.*, 2016).

[3] A relic of a dual appears in some words relating to body parts, e.g. *divskouarn* '[a pair of] ears'.

[4] Often referred to in the literature as '*inflected* or *pronominal* prepositions'.

[5] The Welsh construction using the verb *cael* 'get', e.g. *Cafodd y llyfr ei ddarllen gan Yann* (lit. 'The book got its reading by Yann')

| **Source** | **Description** | **Sentences** | **Tokens** | **Average length** |
|---|---|---:|---:|---:|
| Grammar | Grammar book examples | 277 | 2,092 | 7.55 |
| Bremaik | Magazine articles | 211 | 3,283 | 15.56 |
| OfisPublik | Administrative texts | 177 | 2,119 | 11.97 |
| Wikipedia | Encyclopaedic texts | 136 | 1,935 | 14.23 |
| Examples | Translation examples | 65 | 404 | 6.22 |
| Songs | Traditional songs | 21 | 251 | 11.95 |
| **Total** | | 887 | 10,084 | 11.25 |

**Table 1:** Composition of the syntactically-annotated corpus.

# 4 Corpus

The corpus is composed of texts from a variety of domains (see Table 1 for a breakdown of the composition of the corpus). All of the texts are available under a free/open licence and the resulting corpus is distributed under the terms of the Creative Commons CC-BY-SA licence. In addition to the Breton sentences, each sentence has a translation in French or English. These have been produced either by a human (in the case of the songs, administrative texts and grammar book examples) or by the Breton–French MT system (Tyers, 2010) in the case of the magazine articles and Wikipedia. The texts were chosen to try and cover a range of written domains and grammatical structures. Our final annotated sentences

## 4.1 Preprocessing

Preprocessing the corpus consists of running the text through the Breton morphological analyser[6] available from Apertium (Forcada *et al.*, 2011) and described in Tyers (2009). This analyser also analyses initial consonant mutations[7] performs tokenisation of multi-word units based on the longest match left-to-right. The morphological analyser returns all the possible morphological analyses for each word based on a lexicon of around 18,900 lexemes. After tokenisation and morphological analysis, the text is processed with a constraint-grammar (Bick & Didriksen, 2015) based disambiguator for Breton consisting of 288 rules which remove inappropriate analyses in context. This reduces the average number of analyses per word from around 1.95 to around 1.06.

The native format of the treebank is the VISL format (Bick & Didriksen, 2015). This is a text-based format where surface tokens are on one line, followed by analyses on the subsequent line. The reason for choosing this format was that it was more convenient for hand-annotation, and was the format that the morphological analyser and constraint grammar output. We apply a number of deterministic transformations to convert the VISL format to CoNLL-U and a longest-match set overlap algorithm to convert the tagset from the Breton-specific one to Universal Dependencies.[8]

[6] `https://svn.code.sf.net/p/apertium/svn/languages/apertium-bre`

[7] Initial consonant mutations are where a word changes the first consonant due to morphological or syntactic context, e.g. *kazh* ‘cat’, but *he c’hazh* ‘her cat’.

[8] The conversion code will be released along with the treebank in the final version.

# 5 Annotation guidelines

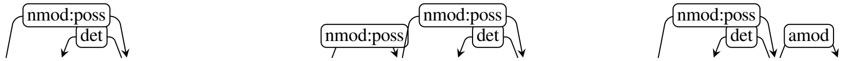
The annotation guidelines are based on Universal Dependencies (Nivre *et al.*, 2016), an international collaborative project to make cross-linguistically consistent treebanks available for a wide variety of languages. The Breton treebank is based on version 2.0 of the guidelines which were published in December, 2016. We chose the UD scheme for the annotation as it provides ready-made recommendations on which to base annotation guidelines. This reduces the amount of time needed to develop bespoke annotation guidelines for a given language; where the existing *universal* guidelines are adequate, they can be imported wholesale into the language-specific guidelines.

The treebank was annotated by a single human annotator; a translation was provided to aid the annotation process.

In the following subsections we describe some particular features of Breton that are interesting or novel with respect to the Universal Dependencies annotation scheme.

## 5.1 Nominal possessive construction

The possessive construction in Breton consists of a juxtaposition of a determined noun phrase with another noun phrase in the form where the non-determined noun phrase comes first (1a). Where there are more than two nouns, the determiner only comes between the two last nouns (1b). Both definite (1a, 1b) and indefinite (1c) determiners may fill the slot.



(1) a. ti ar maer
`NOUN DET NOUN`
*house the mayor*
'the mayor's house'

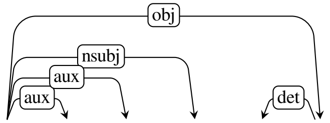
b. toenn ti ar maer
`NOUN NOUN DET NOUN`
*roof house the mayor*
'the roof of the mayor's house'

c. toenn un ti bras
`NOUN DET NOUN ADJ`
*roof a house big*
'the roof of a big house'.

The determiner slot can be also filled with possessive (e.g. *bugale* <u>*ma*</u> *c'hoar* 'my sister's children') and other determiners (e.g. *toenn* <u>*pep*</u> *ti* 'the roof of every house'). The use of `nmod:poss` is quite common across a variety of Universal Dependencies treebanks, including Irish, English and Persian.

## 5.2 Auxiliary items of verbs

There are two main auxiliary items of verbs in Breton, auxiliary verbs, such as *bezañ* 'to be' and *ober* 'to do'. These are used in forming the analytical tenses, such as the present and past (2a).

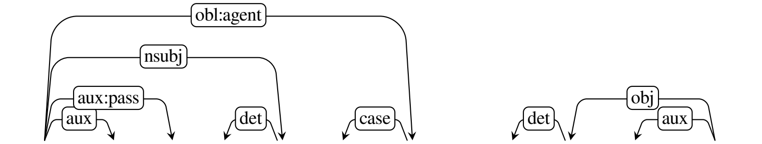


(2) a. Lenn a reas Lenaig al levr
`VERB PART AUX PROPN DET NOUN`
*Read did Lenaig the book*
'Lenaig read the book'.

In common with other Celtic languages, Breton also has a number of verbal *particles* which serve a number of functions including negation and subordination. Both these and the auxiliary verbs are attached as dependents of the main verb (2a) with the relation `aux`. The use of this relation to mark auxiliary verbs is standard UD practice. The attachment of the particles to the main verb as opposed to the finite auxiliary may appear controversial. Grammars of Breton make the verbal particle subordinate to the finite verb. However as auxiliaries in UD may not have dependents, this leaves us with attaching these particles to the main verb.

## 5.3 Passive and impersonal

Unlike other Celtic languages, Breton has an analytic passive, made of the verb *bezañ* 'be' and the past participle of a transitive verb (3a). The agent in this construction may be omitted or expressed with a prepositional phrase using *gant* 'with'. We use the language specific relations `aux:pass` to mark the passive auxiliary and `obl:agent` to mark the demoted agent. Both subtypes are widely used cross-linguistically.



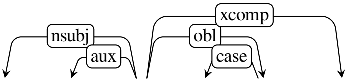
(3) a. Lennet e oa al levr gant Lenaig
VERB PART AUX DET NOUN ADP PROPN
*read was the book by Lenaig*
'The book was read by Lenaig'

b. Al levr a lenner.
DET NOUN PART VERB
*the book was read*
'The book was read.'

Breton also has an automonous (or impersonal) verbal form (3b), like in the other Celtic languages. In this construction the demoted agent cannot be expressed. In the Irish UD treebank (Lynn & Foster, 2016) the core arguments of these verbs are marked with the `obj` and we follow the same convention.

## 5.4 Contracted prepositions

Contractions of prepositions and pronouns (similar to the Spanish *contigo* 'with you') are widespread in the Celtic languages.[9] Unlike the Irish treebank (Lynn & Foster, 2016), which puts features indicating the person information of the contracted pronoun on the preposition, we use the two-level tokenisation scheme of UD to split them into a prepositional part and a pronominal part. Consider the sentence *Petra 'ra dit gouelañ ?* 'What is making you cry ?' (lit. 'What makes to-you crying?') in (4a), the contraction *da* + *it* = *dit* is split into a preposition *d-* and a pronoun *-it*.



(4) a. Petra ' ra d- -it gouelañ ?
PRON PART VERB ADP PRON VERB
*What makes to you cry*
'What is making you cry?'

[9] In the literature they are often called *inflected prepositions*, *conjugated prepositions*, *pronominal prepositions* or *prepositional pronouns*. As they do not have any special syntax we prefer the more cross-linguistic description of *contracted prepositions*.

| System | Lemma | POS | Morph | UAS | LAS |
|---|---|---|---|---|---|
| Maltparser | - | - | - | 73.80 | 65.82 |
| BiST (MST) | - | - | - | 74.78 | 67.00 |
| UDPipe | 88.24 | 89.21 | 87.02 | 75.24 | 68.14 |
| UDPipe [+dict] | 95.82 | 95.24 | 95.12 | 80.29 | 74.71 |
| UDPipe [+dict,+embed] | 95.80 | 95.15 | 95.00 | 79.85 | 74.29 |

**Table 2:** Performance of the systems on a number of tasks. Figures in brackets indicate equivalent scores on UD English, where available.

# 6 Experiments

## 6.1 Parsing performance

In order to test the treebank in a real setting, we evaluated three widely-used popular dependency parsers: Maltparser (Nivre *et al.*, 2007), UDPipe (Straka *et al.*, 2016) and BiST (Kiperwasser & Goldberg, 2016). In addition we provide results for using the treebank for part-of-speech tagging using UDPipe.

For Maltparser we used the default settings and for BiST parser we tested the MST algorithm.

We performed 10-fold cross-validation by randomising the order of sentences in the corpus, and splitting them into 10 equally-sized parts; in each iteration, we held out one part (90%) for testing (89 sentences) and used the rest (10%) for training (801 sentences). As the BiST parser required additional heldout data, we performed a 80-10-10 training-dev-test split per iteration. We calculated the labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for each of the models using the CoNLL-2017 official evaluation script.[10] The same cross-validation splits were used for training all three parsers.

The morphological analyser and part-of-speech tagger in UDPipe was tested both with and without an external morphological dictionary. In this case the morphological dictionary, shown in Table 2 as `[+dict]`, consisted of a full-form list generated from the morphological analyser described in §4.1 numbering 296,905 entries. Further, the analyser was trained with dimension 300 fastText embeddings (Bojanowski *et al.*, 2016) `[+embed]`. These, unfortunately, did not improve our parsing results; we hypothesise that this was due to poor tokenisation of the embeddings training corpus (a generic 'language-independent' tokeniser will likely treat the Breton *c'h* letter incorrectly), and propose experiments on alternative forms of tokenisation for future work. Full results are presented in Table 2. Lemma, POS and morphology scores are absent for Maltparser and BiST as they are not included; each was evaluated on the output of a UDPipe instance (without embeddings and a dictionary). For comparison, a similarly set-up UDPipe instance (without external dictionaries or embeddings) achieves an LAS of 77.25 on English, 80.50 on French and 62.87 on Irish, which is likely the most comparable to our Breton treebank (Straka & Straková, 2017).

[10] `http://universaldependencies.org/conll17/evaluation.html`

# 7 Future work

The most obvious avenue for future work is to annotate more sentences. A treebank of 10,000 tokens is useful — it can be used for bootstrapping and also is key for evaluating unsupervised or semi-supervised systems — but in order to be able to train a parser useful for parsing unseen sentences we would need to increase the number of tokens 6–10-fold.

There are a number of quirks in the conversion process from VISL to CoNLL-U, for example the language-independent longest-common-subsequence algorithm could be replaced with a Breton-specific one that would be able to successfully split tokens like *en* (when it stands for 'in the') into *e* and *n* — the current generic algorithm gives *en* and *n*. We are also interested in collaborating with the authors of the Irish treebank to improve cross-linguistic compatibility.

# 8 Concluding remarks

We have described the first syntactically-annotated corpus of Breton. The treebank will be used as one of the languages in the 2018 CoNLL on dependency parsing and has been released for public use.[11] The corpus consists of a little over 10,000 tokens and is released under a free/open-source licence.

# References

BICK E. & DIDRIKSEN T. (2015). Cg-3 – beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA*, p. 31–39: Linköping University Electronic Press, Linköpings universitet.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

FORCADA M. L., GINESTÍ-ROSELL M., NORDFALK J., O'REGAN J., ORTIZ-ROJAS S., PÉREZ-ORTIZ J. A., SÁNCHEZ-MARTÍNEZ F., RAMÍREZ-SÁNCHEZ G. & TYERS F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, **25**(2), 127–144.

HEMON R. (2007). *Breton Grammar*. Evertype, 2nd edition.

KIPERWASSER E. & GOLDBERG Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *TACL*, **4**, 313–327.

LYNN T. *et al.* (2016). Irish dependency treebanking and parsing.

LYNN T. & FOSTER J. (2016). Universal Dependencies for Irish. In *Proceedings of CLTW 2016*.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIČ J., MANNING C., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference (LREC'16)*.

NIVRE J., HALL J., NILSSON J., CHANEV A., ERYIGIT G., KÜBLER S., MARINOV S. & MARSI E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95–135.

[11] https://github.com/UniversalDependencies/UD_Breton-KEB

Poibeau T. (2014). Processing mutations in Breton with finite-state transducers. In *Proceedings of the Celtic language technology workshop (CLTW) organised with COLING2014*.

Press I. J. (1986). *A Grammar of Modern Breton*. Mouton Grammar Library. Mouton.

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

Straka M., Hajič J. & Straková J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France: European Language Resources Association (ELRA).

Straka M. & Straková J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada: Association for Computational Linguistics.

Ternes E. (2008). Breton. In D. MacAulay, Ed., *The Celtic Languages*, Cambridge Language Surveys. Cambridge University Press, 1 edition.

Tyers F. M. (2009). Rule-based augmentation of training data in Breton–French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*, p. 213–218.

Tyers F. M. (2010). Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation, EAMT10*, p. 174–181.

Zeman D., Popel M., Straka M., Hajic J., Nivre J., Ginter F., Luotolahti J., Pyysalo S., Petrov S., Potthast M., Tyers F., Badmaeva E., Gökirmak M., Nedoluzhko A., Cinkova S., Hajic jr. J., Hlavacova J., Kettnerová V., Uresova Z., Kanerva J., Ojala S., Missilä A., Manning C. D., Schuster S., Reddy S., Taji D., Habash N., Leung H., de Marneffe M.-C., Sanguinetti M., Simi M., Kanayama H., dePaiva V., Droganova K., Martínez Alonso H., Çöltekin c., Sulubacak U., Uszkoreit H., Macketanz V., Burchardt A., Harris K., Marheinecke K., Rehm G., Kayadelen T., Attia M., Elkahky A., Yu Z., Pitler E., Lertpradit S., Mandl M., Kirchner J., Alcalde H. F., Strnadová J., Banerjee E., Manurung R., Stella A., Shimada A., Kwak S., Mendonca G., Lando T., Nitisaroj R. & Li J. (2017). Conll 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 1–19, Vancouver, Canada: Association for Computational Linguistics.