



Old Catalan Morphosyntax: Developing an Annotated Corpus

MARIEKE MEELEN 

AFRA PUJOL I CAMPENY 

**Author affiliations can be found in the back matter of this article*

RESEARCH PAPER

]u[ubiquity press

ABSTRACT

This paper presents a full procedure for the development of a Part-of-Speech (POS) tagged corpus of Old Catalan. As an extremely low-resource language with rich inflection and frequent homographs, Old Catalan poses non-trivial problems in the development of a searchable constituency-based treebank. We demonstrate, however, that a semi-supervised method of incrementally building training data using both neural and memory-based taggers, together with the Pyrrha annotation tool is highly efficient and yields accurate results. We propose that this simple and effective method could easily be extended to other low-resource historical languages for which no NLP tools exist yet.

CORRESPONDING AUTHOR:

Marieke Meelen

University of Cambridge,
Cambridge, United Kingdom

mm986@cam.ac.uk

KEYWORDS:

Old Catalan; POS tagging;
historical treebank

TO CITE THIS ARTICLE:

Meelen, M., & Pujol i Campeny, A. (2021). Old Catalan Morphosyntax: Developing an Annotated Corpus. *Journal of Open Humanities Data*, 7: 30, pp. 1–12. DOI: <https://doi.org/10.5334/johd.54>

Catalan is a Romance language that originated to the east of the Pyrenees mountain ridge during the Middle Ages. Its affiliation to either the Ibero-Romance or the Gallo-Romance groups within the Romance languages has been debated since the emergence of historical Romance linguistics, and is still contended. Catalan possesses a Medieval textual record that, while being limited, offers us enough data to allow for an in-depth study of this language. To this date, the largest publicly available searchable corpus of Old Catalan is the Corpus Informatitzat del Català Antic, the ‘Digitised Corpus of Old Catalan,’ commonly referred to as CICA (Torruella et al., 2009).¹ CICA contains 414 texts dating from the 11th to the 18th century, and it allows for simple and complex token and lemma searches, enabling the study of collocations or the distribution of specific lexical items. Nevertheless, as the texts contained in CICA are not morphosyntactically annotated, this corpus does not lend itself to the study of the morphosyntax of Catalan diachronically. The lack of an open-access fully parsed corpus or treebank of Old Catalan renders it inaccessible to the research community for certain types of linguistic and philological studies, such as those focusing on diachronic syntax, information structure and patterns of language change. With the ultimate objective of developing a morphosyntactically annotated corpus of Old Catalan, here we present a Part-of-Speech (POS) tagger developed for Old Catalan and trained with a 13th century chronicle: *Llibre dels Fets* (composed between 1229 and 1276). This text has been deemed suitable for the training of the POS tagger, and in the future, syntactic annotator, on the basis of three factors:

- i. It is the first *crònica* ‘chronicle’: The *Llibre dels Fets* is the first of the so-called Great Catalan Chronicles, four historiographical texts written in prose from the end of the 13th century and throughout the 14th. While they differ in form, they all have a common theme: they narrate and praise the feats of several Catalan kings who reigned during the 13th and 14th centuries in the Crown of Aragon, a Medieval kingdom that resulted from the dynastic union of the Kingdom of Aragon with the County of Barcelona and its vassal territories. *Llibre dels Fets* broke with the preceding historiographic tradition modelled on Latin *annales* (lists of dated historical events) by narrating the feats of King James I in the 1st person, using a vivid style that has been described as ‘spontaneous, colloquial, primitive and careless’ (Bruguera, 1991) due to the abundance of direct reported speech, code-switching into languages other than Catalan, references to the book’s audience² and the presence of information of personal nature, among other traits. Koch & Oesterreicher (2012) propose assessing linguistic variation along the immediacy/distance axis, instead of the written/oral one. Within their framework, the *Llibre dels Fets* would be a written text with features associated with immediate language, closer to the language of informal spoken interactions linked to informal registers (Bieber & Conrad, 2009).³ Given that the text does not abide by rigid literary conventions which, in Medieval times, often favoured the use of Latinising syntax, and that it exhibits a high degree of oral-like features, it provides us with a unique opportunity to investigate the word order of 13th century Old Catalan.
- ii. The probable involvement of King James I in the production of the text: The Medieval notion of *authorship* and *author* differ significantly from the modern ones, where literary works are attributed to the individual or collective that has produced them and cannot be lawfully altered without explicit permission from the author/s. In Medieval times, literary works were not seen as immutable entities. They could (or not) be attributed to

1 Another tool that allows for form and lemma search in context is the *Diccionari de Textos Catalans Antics* (Dictionary of Old Catalan Texts, henceforth DTCA) (Rafel i Fontanals, 2009). This corpus currently contains 25 works dating from the 11th to the 15th century that have been lemmatised and morphologically annotated, but not syntactically. Therefore, one cannot extract results on the basis of the syntactic position of a token. Additionally, the DTCA does not allow for the search of collocations, which CICA does. Since DTCA does not allow for the search of collocations that might reflect specific syntactic structures, it is not well suited for the study of word order phenomena at large scale.

2 *Llibre dels Fets*, like many other medieval literary works, was conceived to be read aloud to an audience. This is both due to the low level of literacy during the period, even in higher circles, as well as to the complexity and cost of manuscript production, which prevented large scale reproduction and circulation of texts.

3 We are aware that this corpus, at present, instantiates only a very specific register of Old Catalan, and thus, it has limitations as to the conclusions that can be drawn from its data, as is often the case with corpora that compile texts belonging to one register only (Kytö & Smitterberg, 2015). In the future, we plan to expand it with texts contemporary to *Llibre dels Fets* in order to make it representative of the Old Catalan written record.

someone, but each scribe or reader was free to intervene in the text and add or subtract any material as they saw fit without *creating* a new text.⁴ Therefore, it is not possible to say that James I is the author of *Llibre dels Fets* in the modern sense. However, given the abundance of oral-like traits (Soldevila, 1971; Bruguera, 1991; Pujol i Campeny, 2021) and intimate information found throughout the text, it is possible that he dictated most of its content, and that the text was later put together as a Chronicle by a scribe.⁵ If the text was indeed dictated by the King, this would bring us closer to the modern notion of authorship and we could take the text to represent his speech without having undergone too much change. Regardless of whether the King did actually utter the words found in the text or not, we do know that the different manuscripts of *Llibre dels Fets* that have reached our day show discrepancies in terms of spelling or choice of lexical items stemming from scribal intervention while they remain stable in terms of word order.

iii. A linguistically-aware edition of the text is available: Medieval Catalan texts present editors with challenges due to the lack of standardisation of the orthography and the presence of scribal errors which render certain fragments of a text difficult to understand. Bruguera's (1991) edition of *Llibre dels Fets* takes a linguistically-aware perspective in producing a text that relies on the oldest manuscript that has reached our days (manuscript H, 1343) to produce a regularised (not standardised) version of the text that (i) spells out abbreviations; (ii) regularises upper and lower case letters; (iii) regularises word separation, accentuation and punctuation according to standard Modern Catalan rules (elisions that are not currently rendered graphically in standard Modern Catalan are marked with a *punt volat* 'flying dot', '·'); (iv) it marks both thematic chapters established by the editorial tradition of the text, while also keeping the folia annotation. Given the qualities of this edition, it is a good starting point to produce an annotated version of the text, as abbreviations have already been spelled out. At the same time words are conveniently separated, while preserving greatly valuable spelling particularities which can point towards language change. Bruguera's edited text was updated for its inclusion in CICA, with the addition of graphic accents contributing to the disambiguation of homographs. We have worked with this version of Bruguera's edition.

2 ESTABLISHING THE MORPHOSYNTACTIC TAG SET

The investigation of the evolution of syntactic constructions and the distribution of different forms diachronically benefits heavily from the analysis of large corpora that are consistently annotated. Manual annotation of thousands (and ideally, millions) of words is an extremely time-consuming task that is very prone to error. The use of Natural Language Processing (NLP) tools to automatise the annotation process and consistently treat large amounts of data in a minimal period of time have been successfully applied to various languages for the production of historical corpora, including Welsh, English and French.

Like most Medieval languages, Old Catalan had not undergone standardisation, and therefore, it displayed a great degree of variation at orthographic level. While the regularisation of spelling is possible, it is a time-consuming task that requires editorial decisions of philological nature in order to establish which form should be used as the 'standard' or 'regularised' form. In addition, the standardisation of the text can conceal dialectal variation as well as differences in scribal practices, impoverishing the text for the purposes of the diachronic study of a language. In the specific case of *Llibre dels Fets*, it would have concealed features that have attracted the interest of many researchers, as Colón Domènech's (2012) volume shows.

One of the key steps in the development of a POS tagger is the definition of the tag-set, since it interacts directly with the effectiveness of the POS tagger (a lower number of tags makes morphosyntactic classification easier, thus yielding better results) and it potentially limits the possible searches that can be carried out. For the Old Catalan POS tagger, the defined tag-set

⁴ For more on the notion of author and authorship in medieval Europe, see Partridge & Kwakkel (2012).

⁵ The *Llibre dels Fets*'s composition probably began during the conquest of Mallorca, in 1224, and it ended with King James I's death, in 1276. The preface and the ending of the book were probably added after the King's death by another author, which explains the differences in tone (more erudite) and use of Latin (Soldevila, 1971; Vinas & Vinas, 2008).

is based on the standard UPenn annotation scheme,⁶ in order to render it readily comparable with similar resources in the UPenn historical corpora collections. At the same time, it has been simplified where possible and enriched where needed in order to adapt it to Old Catalan grammar and to provide as much information as possible for the study of word order variation at clausal level. As a result, tags devoted to the nominal domain have been simplified, as Old Catalan does not systematically display comparative and superlative morphology for adjectives, for instance, and different determiner categories have been created in order to reflect the emerging article system that distinguishes proper names from other nouns.

Certain part of speech categories have been further specified with different grammatical attributes, such as case, person and number. These are added to the part of speech label separated by a delimiter ^. Therefore, a third person plural accusative pronoun would be tagged as PRO^A^3^PL, with the attributes case, person and number occurring in this order, following the convention developed for the HeliPaD corpus (Walkden, 2016). A comprehensive list of the tags and attributes used in this corpus can be found in the Appendix. Our fully-tagged corpus is deposited and made available open access on Zenodo (<https://doi.org/10.5281/zenodo.5615759>), as are the word embeddings we created for the neural-based tagger (<https://doi.org/10.5281/zenodo.5615556>).

2.1 VERBAL TAGS AND CHALLENGES WITHIN THE VERBAL DOMAIN

As is the case in most Romance languages, verbal inflection is expressed through suffixes attached to the verbal root. Inflected verbs are specified as follows: firstly, the word category tag VB appears, followed by tense (P for present, D for past, F for future,) and, in turn, followed by mood (I for indicative, S for subjunctive and C for conditional⁷). The verbal tag is then completed with person and number information (^1/^2/^3 + ^SG/^PL). Therefore, a verb tagged VBPI^1^SG would indicate a verb in the present indicative, first person singular. In order to keep tags to a minimum, no aspectual tags were added, as most perfect tenses are expressed through verbal periphrases (auxiliary + non-finite form,) with the exception of the synthetic past perfect indicative. There are no tags for passive morphology either, as it was also expressed by means of analytic constructions.

Past perfect participles can agree with an element from their context in gender and number. While they were not tagged for these categories, whether they are inflected and agree with an element from their context is specified with the addition of the tag I ('inflected') to the past perfect participle label VN, yielding VNI.

Within the Catalan verbal domain we find a rich paradigm of pronouns that cliticise onto verbs and pose a real challenge for the tagger. The clitic pronoun paradigm includes reflexive, accusative and dative clitic pronouns that distinguish person and number (and in the case of 3rd person accusative pronouns, also gender: masculine, feminine and neuter). The paradigm also counts with two adverbial pronouns.⁸

The paradigm exhibits a high degree of syncretism between reflexive, accusative and dative clitic pronouns in the 1st and 2nd persons. This is illustrated in examples (1–3) using the 1st person reflexive, accusative and dative clitic pronoun *em*, *me*, *m'*:

- (1) *Primerament vos dich que no m'acort a re que*
 firstly you-DAT say-1SG that not I-1SG.RFL=agree-1SG to anything that
vosaltres digats (...)
 you say-2PL
 'Firstly, I tell you that I do not agree with anything that you say'

Fol. 151r, l. 8

⁶ See <http://www.ling.upenn.edu/histcorpora>.

⁷ The imperative mood tag I appears attached immediately after the VB label, yielding VBI. Since in Old Catalan the imperative mood does not express tense, this category is not specified.

⁸ See §18 in the *Gramàtica de la Llengua Catalana* (Institut d'Estudis Catalans, 2016) and Colomina i Castanyer (2008) for more on the Modern Catalan clitic pronoun paradigm and Batllori et al. (2005) and Fischer (2011) for more on the Old Catalan clitic paradigm.

- (2) *e prec-vos que levets m'í, (...)*
and beg-1SG=you-DAT that take-2PL.SBJV =me=there-LOC
'and I beg you to take me there, (...)'

Fol. 143v, l. 12

- (3) *car ells m'ó àn dit (...)*
since they me-DAT=it-ACC have-3PL said
'since they have told it to me (...)'

Fol. 125r, l. 18

Etymologically, third person accusative clitic pronouns derive from Latin demonstratives *ILLE ILLA ILLUD*, which are also the origin of one of the sets of definite articles found in the language (the other being derived from the demonstrative *IPSE IPSA IPSUD*). This accounts for the homophony between definite articles and 3rd person accusative clitic pronouns. This is the case of the masculine plural article *los* (see 4,) the 3rd person plural masculine accusative clitic pronoun *los* (see 5,) which, additionally, are also homophonous with the 3rd person plural dative clitic pronoun *los* (see 6).

- (4) *los infants e la gent menuda hagren gran alegria (...)*
the children and the people little had-3PL great happiness
'the children and lay people were joyful (...)'

Fol. 17v, l. 22

- (5) (a) *E mostraren -los -nos (...)*
and showed-3PL =them-ACC =we-DAT
'And they showed them to us'

Fol. 17r, l. 4

- (b) *e mataven e enderrocaven dels moros ali on los trobaven.*
and killed-3PL and destroyed-3PL of;the moors there where them-ACC =found-3PL
'and they killed and destroyed anything belonging to the moors wherever they found them.'

Fol. 36r, l. 11

- (6) (a) *E dixem-los que (...)*
and said-1PL=them-DAT that
'and we told them that (...)'

Fol. 197v, l. 4

- (b) *Encara los dixem que (...)*
still them-DAT= said-1PL that
'We even said to them (...)'

Fol. 62v, l. 22

Old Catalan had two adverbial clitic pronouns: *en* and *hi*. The tagging of pronominal *en* is especially challenging, given its homophony with the atonic preposition *en* and the title *En* applied to some masculine names (akin to English 'Sir,') as illustrated in examples (7-9):

- (7) *Lexats-ho sobre nós que a la longa nós vos en guardarem*
leave-2PL.IMP=it-ACC on we that at the long we you-DAT of;it= keep-1PL.FUT
d'onta e de vergonya.
of;shame and of embarrassment
'Leave it to us because in the long run we will keep you from shame and embarrassment'

Fol. 122r, l. 1

- (8) *can foren en Catalunya (...)*
when were-3PL in Catalonia
'when they arrived in Catalonia'

Fol. 5v, l. 19

- (9) *E, quan fo presa València, vench En Ramon Folch de Cardona, e (...)*
and when was-3_{SG} taken-FEM Valencia, came-3_{SG} Sir Ramon Folch de Cardona, and (...)
'And, once Valencia was taken, Sir Ramon Folch of Cardona came, and (...)'
Fol. 122r, l. 11

So far we have identified synchretism within the clitic pronoun paradigm and homophony between clitic pronouns and other lexical items as major challenges for the tagger within the verbal domain. In addition, we encountered other challenges that required special attention at the manual correction stage, stemming from the high frequency of homophonous (and homographic) forms across tenses and persons within the Old Catalan verbal paradigm. Amongst them, we find⁹: homophony between the 1st and 3rd person singular of the past imperfect indicative (example 10), homophony between the 3rd person singular of the past perfect of certain verbs with the past participle of the same verb, as is the case of *promès* (example 11), among others.

- (10) (a) *E ço que yo feya, (...)*
and that which I did-1_{SG}
'And what I did, (...)'
Fol. 20r, l. 2
- (b) *e, a l'entrar que él feya en la tenda, anam-lo pendre als*
and at the=entering that he did-3_{SG} in the tent went-1_{PL}=him take-INF by;the
cabeyls e traguem-lo.n.
hair and took-1_{PL}=him=from;it
'and, as he entered the tent, we took him by the hair and dragged him out.'
Fol. 132v, l. 14
- (11) (a) *e promès-nos complir tot açò damunt dit.*
and promised-3_{SG}=US-DAT accomplish-INF all that above said
'and he promised to accomplish all the aforementioned things.'
Fol. 20r, l. 2
- (b) *faent aquel compliment que-ns havia promès, (...)*
doing that commitment that;US-DAT= had-3_{SG} promised
'abiding by the commitment that he had made to us, (...)'
Fol. 75v, l. 7

2.2 NOMINAL TAGS AND CHALLENGES WITHIN THE NOMINAL DOMAIN

Within the nominal domain, tags were kept to a minimum. Nouns, adjectives and determiners were labelled for their category (N for nouns, ADJ for adjectives, D for determiners,) but not for gender or number.

Proper person names could be preceded by the proper noun articles *Don/En* for masculine nouns, and *Dona* for feminine nouns. The category DPR was created to account for them and distinguish them from other determiners.

Some Catalan pronouns and determiners (labelled D) are indefinite quantifiers. Their distribution exhibits particularities when compared to that of other members of the determiner class and, therefore, it was in our interest to mark them differently. Since they do not form a natural part of speech category, they have been further specified with the attribute D^Q.¹⁰

Tonic pronouns are marked for person and number, in contrast with clitic pronouns, which receive further specification (see section 2.1).

⁹ This is by no means an exhaustive list of the different homographs present in the Old Catalan textual record. Another challenge that we will not explore at length here in the interest of space is the problematic posed by the analytic future and conditional tenses, which consist of an infinitive and an auxiliary homophonous (and therefore homograph) with the verb *haver* 'to have', separated by a clitic pronoun. Our proposed solution here is to tag each element separately, in order to enable searches of **infinitive-clitic pronoun-inflected verb 'to have'**.

¹⁰ For instance, a particularity of quantified phrases is that they can be moved from their canonical position to receive an emphatic reading elsewhere in the clause. For a more nuanced analysis of the distribution of quantified phrases and word order in Old Catalan see Pujol i Campeny (2018).

One of the challenges that we encountered regarding the tagging of nominal categories is the abundance of non-finite verbal forms that underwent nominalisation and thus, have a homophonous nominal counterpart. These mainly include past participles (i.e., *vinguda, presa, feyt, anada, estada, partida*, meaning ‘arrival, taking, event/fact, coming, stay, leaving, example 12) and infinitives (i.e., *poder, saber*, ‘to be able to, to know’, example 13).

(12) (a) *e nós que fariem gran nostre pro e gran nostra honor ab la*
and we that do-3_{PL.COND} great our profit and great our honour with the
lur venguda.
their coming
‘and we [said] that we would increase our fortune and our honour with their coming.’
Fol. 148r, l. 18

(b) *E quan ella fo venguda, (...)*
and when she was-3_{SG} come-PPT
‘And when she arrived, (...)’
Fol. 20r, l. 19

(13) (a) *cuydaven enganar tots los alters ab lur saber, (...)*
tried-3_{PL} deceive-INF all the others with their knowledge
‘they tried to deceive all the others with their knowledge.’
Fol. 56r, l. 23

(b) *E nós volguem saber dels altres, si eren en aquel consel (...)*
and we wanted-1_{PL} know-INF of;the others if were-3_{PL} in that advice
‘And we wanted to know about the others, if they agreed with that piece of
advice, (...)’
Fol. 104v, l. 21

Finally, proper names presented a challenge, as most of them constitute unseen words and present unpredictable morphology. However, after the third round of training, the accuracy in which the tagger successfully identified them increased significantly, most likely due to the tagger being able to recognise capital letters.

3 POS TAGGING HISTORICAL LOW-RESOURCE LANGUAGES FROM SCRATCH

For extremely low-resource languages with complex morphology and a large amount of short, homophonous forms like Old Catalan, even basic NLP tasks like part-of-speech tagging can prove challenging at first. Apart from the digitised edition of the text, no further resources were available to us for Old Catalan at this point.

In order to train the tagger, we therefore first manually annotated a text sample containing 4,500 words. Manual annotation was carried out through with the Pyrrha annotation tool (Clérice et al., 2021). Pyrrha allows for manual annotation of lemmas, morphological features and POS tags. Legitimate POS tags can be entered as so-called ‘control lists’ ahead of tagging. These options are then made available by means of a dropdown list while annotating, thus avoiding mistakes that could easily occur when typing every POS label separately. Pyrrha can furthermore automatically extrapolate from already tagged tokens: identical and/or similar tokens in the rest of the corpus can be presented in a convenient list of instances in context and their POS tags can then be adjusted in bulk: it allows annotators to decide whether the same tag is to be applied to all similar tokens, to only some, or to none. While this feature significantly speeds the tagging process, manual annotation remains a time-consuming endeavour. We therefore decided to limit this initial time-consuming round of manual annotation to 4,500 tokens. While getting used to our newly designed tag set, this initial task took 32 working hours (140,6 words/hour). In order to create a sizeable gold standard, we then proceeded in a semi-supervised manner, incrementally building a large training set using memory-based and neural taggers.

3.1 INCREMENTALLY BUILDING UP OUR TRAINING SET

The manually annotated 4,500 tokens allowed for the training of the memory-based POS tagger (MBT based on TiMBL¹¹). After each round of manual annotation or correction, we used the MBT to generate a new tagger based on the growing set of training data. We then use this new tagger to tag the rest of the corpus, as we predicted that the increasing accuracy rates would make subsequent correction less time-consuming.

The second round of manual correction resulted in a new training set of 10,000 tokens that were corrected over 32 working hours (312,5 words/hour). In turn, a further 10,000 tokens were manually corrected reaching a total of 20,000 corrected tokens over 18 hours, at the rate of 556 words/hour. With the third round of manual correction, we reached 40,000 corrected tokens over 23h30min, at 851 words/hour. In the final round, a further 20,000 tokens were manually corrected, at a 1666.7 words/hour rate. **Table 1** shows how much time was invested in manual annotation and correction in this semi-supervised, incremental build-up of the training data.

TRAINING ROUND	HOURS INVESTED	WORDS/HOUR	WORD/HOUR INCREASE
0-4,500	32	140,6	0%
4,500-10,000	32	312,5	122,26%
10,000-20,000	18	555,6	295,16%
20,000-40,000	23.5	851,6	505.59%
40,000-60,000	12	1666.7	1085.42%

Table 1 Incremental build-up and time investment.

From **Table 1** it is clear that even though this POS tagger was initially trained with a very small sample of 4,500 tokens only, our semi-supervised method renders the process of tagging almost three times faster than through manual annotation of larger data sets from the outset. The high increase in correction rates are partly due to increased experience of the annotator (including familiarity with the data and tag set,) but also due to the efficiency of working with Pyrrha, which allows for automatic extension of POS labels to similar tokens. Most importantly, however, the global accuracy of the memory-based tagger increased significantly with each round of training; we present the parameters and further technical details along with the results in the following section.

3.2 MEMORY-BASED VS NEURAL TAGGING

We initially chose the TiMBL's memory-based tagger (MBT,) because it has yielded very good results for historical, low-resource languages in the past (see, for example, the results for the Middle Welsh corpus in Meelen (2016) or the Tibetan historical corpora in Meelen et al. (2021)). In recent years, however, POS taggers based on neural networks have yielded very good results. One particularly good off-the-shelf model is TARGER, a BiLSTM-CNN-CRF tagger by Chernodub et al. (2019).¹² Neural taggers generally work well on large data sets. As our initial set of training data was extremely small, we skipped the first round and only started testing once we had a gold standard of 10,000 tokens. Word embeddings, which are essential for TARGER to perform well, were created with FastText (www.fasttext.cc). The total number of tokens of Old Catalan material available to us at present was with just over 156k tokens much too small to create very good embeddings, but at least it could facilitate neural tagging. We evaluated the results of both 10k and 60k gold standards, each divided into 80/10/10 splits for training, development and test sets with the following key parameter settings:

```
batch_size=10
char_cnn_filter_num=30
char_embeddings_dim=25
```

¹¹ The MBT is freely available here: <http://languagemachines.github.io/mbt/>.

¹² TARGER is freely available here: <https://github.com/achernodub/targer>.


```
char_window_size=3
check_for_lowercase=True
dropout_ratio=0.5
epoch_num=100
lr=0.01
lr_decay=0.05
min_epoch_num=50
model='BiRNNCNCRF'
rnn_hidden_dim=100
rnn_type='LSTM'
```

TOKENS	GLOBAL ACCURACY
10,000	85.5
60,000	91.4

Table 2 TARGER results.

A global accuracy of 91.4% is not bad considering the large tag set consisting of 114 distinct POS labels. However, results could most certainly be improved once more Old Catalan data becomes available so that better word embeddings can be created. In addition, higher accuracies could be achieved through feature engineering and by switching from a recurrent to a convolutional neural network or by adding further (Bi)LSTM and Conditional Random Field layers. We leave this for future research.

Unlike TARGER, the memory-based tagger (MBT) does not need large data sets or word embeddings to get decent results on challenging historical data like our Old Catalan corpus. The MBT allows for different parameter settings according to features of the words themselves or the context in which they appear. We started testing the default settings, but then adjusted the parameters for known and unknown words so that morphological suffixes in particular could feed better into the morphosyntactic classifier. We specifically focused on context, selecting the maximal windows for tags preceding (d) and following (a) the focus word (f). In addition, for unknown tokens, we made the tagger focus on the last three characters (s) in order to make optimal use of morphological suffixes in Old Catalan, which usually consist of suffixes containing up to 3 letters. The optimal settings tested so far for Old Catalan therefore are (see the MBT manual for further details Daelemans et al. (2010)):

```
-p dddfaaa -P sssdddFchnaaa
```

The accuracy of the memory-based POS tagger has also improved with each training round, reaching levels of accuracy akin to those of a human annotator for seen words and close to 96% globally:

TOKENS	GLOBAL ACCURACY	KNOWN WORDS	UNKNOWN WORDS
4,500	86.1	94.6	46.8
10,000	91.3	96.4	56.4
20,000	93.8	97.6	58.4
40,000	94.7	97.2	64.3
60,000	95.8	97.6	68.1

Table 3 MBT results per training round.

Through a 10-fold cross-validation, we calculated the Precision (percentage of system-provided tags that were correct,) Recall (percentage of tags in the input that were correctly identified by the system) and F-score or Global Accuracy (weighted harmonic mean of recall and precision). For the individual categories, Precision and Recall give more insight in the degree to which the

model over- or under-generalises certain tags. When analysing the results in more detail, we see that when it comes to both known and unknown words, frequency of the specific POS tag plays an important role in overall accuracy. For both known and unknown words, there are some tags that occur very infrequently (e.g. 1 or 2 times). These are mostly pronominal clitics, whose tags include specification for morphological case, number and person (see the lower part of [Tables 5](#) and [4](#)). Precision, Recall and F-Scores for those are generally low, not only because they are infrequent, but also because there are a number of homophonous tokens with these POS labels, as we have shown in examples 1–9 above.

POS TAG	PRECISION	RECALL	F-SCORE	N
N	0.74	0.75	0.74	856
NPR	0.9	0.94	0.92	585
VNI	0.66	0.67	0.67	219
VBDI ³ PL	0.73	0.75	0.74	188
VBDI ³ SG	0.65	0.68	0.66	173
PRO ³ PL	0	0	0	1
PRO ^A 3 ³ SG	0	0	0	1
PRO ^D 2 ² PL	0	0	0	1
PRO ^D 3 ³ SG	0	0	0	0
OLB	0	0	0	0

Table 4 MBT results for highest and lowest frequency unseen tokens.

POS TAG	PRECISION	RECALL	F-SCORE	N
P	1	1	1	6191
N	0.98	0.98	0.98	6185
CONJ	1	1	1	4919
C	1	1	1	4139
COMMA	1	1	1	3910
PRO ^{RFL} 2 ² SG	1	1	1	2
ADJ ^{POS}	0	0	0	2
VBI ² SG	0	0	0	2
VBDI ² PL	0	0	0	1
”	0	0	0	0

Table 5 MBT results for highest and lowest frequency tokens.

High frequency, on the other hand, unsurprisingly leads to higher accuracies. The parameters of the MBT were set to pay attention to initial capital letter, which clearly yields the high F-scores for proper nouns (NPR). Even when names are unknown yet, the tagger can accurately predict the tag because of the initial capital letters. Other frequent POS tags for unknown tokens are regular nouns (N) and third-person past-tense verb forms (VBDI³SG/PL). Recall results are slightly higher than Precision here, but overall F-scores vary between 64% and 77%. Again, unsurprisingly, the most-frequent tags for known tokens score extremely high, with only nouns not reaching 100% accuracy in the top 5 presented in [Table 5](#) (for a full overview of the MBT results, see Appendix).

Generally, with a global accuracy of 95.8% the MBT trained on 60k manually corrected performs very well. The >1% increase in F-score between the 40k and 60k training sets furthermore suggests it is still possible to get higher accuracies as we keep on extending our training data. The high accuracy rates, in combination with Pyrrha’s efficient manual correction and tag extension function, mean that the time annotators need to spend creating gold standards of the POS-tagged corpus is significantly reduced.

4 CONCLUSION

This paper describes our pipeline to create the first morphosyntactically annotated corpus of Old Catalan. We first presented tag set and annotation manual for Old Catalan, based on that used for the standard UPenn corpora, with specific extensions especially in the domain of case, person, and number features. This resulted in a large number of POS tags (>110 in total,) which presents a real challenge for any automatic morphosyntactic classifier. Without those additional features, however, future research opportunities for scholars in Catalan studies, cross-linguistic syntax and beyond, would be extremely limited. In addition, these extra agreement features will facilitate future conversion of this corpus, which, like the other UPenn historical corpora, will be constituency-based to conll-U dependency formats as well.

Building on previous work on other historical, low-resource languages like Middle Welsh and Classical Tibetan, we presented the results of our semi-supervised method consisting of five iterations of manual tagging and correction, incrementally building up our training set to >60k tokens. After each iteration, the memory-based tagger (MBT) was trained and the entire corpus was retagged based on the newly trained version, thus speeding up manual correction of subsequent batches. In addition to generating a memory-based tagger for Old Catalan, we also created word embeddings with FastText in order to test TARGER, a BiLSTM-CNN-CRF tagger. Because of the small dataset, global accuracies were still lower for TARGER. Once more digitised Old Catalan data will be made available this tagger will most likely yield much higher accuracies as well.

For small and highly complex data sets like our Old Catalan corpus, semi-supervised, incremental annotation methods like these can thus yield highly accurate morphosyntactic taggers with minimum effort. Depending on the complexity of the tag set, multiple iterations of manual correction might be necessary to get started, but with each iteration, results clearly improve and the time invested on subsequent correction sessions is significantly reduced. This semi-supervised method of memory-based tagging, combined with manual correction in Pyrrha, is thus highly efficient to create reliable training data for historical, low-resource and morphologically rich languages.

The data and code for this paper can be found on our GitHub repository: <https://github.com/lothelanor/catalancorpora>.

ACKNOWLEDGEMENTS

Research that partially facilitated the work presented in this article was funded by the British Academy (PDF grant pf170063), and the Cambridge Humanities Research Grant (tier 1 grant, GANT011262). Additionally, this work has been supported by the French government, through the UCAJEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number C870A06228 – EOTP : SYVACA – D112.

ADDITIONAL FILE

The additional file for this article can be found as follows:


- **Appendix.** This appendix contains our annotation manual and a detailed analysis of the results of all tags. DOI: <https://doi.org/10.5334/johd.54.s1>

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Marieke Meelen  orcid.org/0000-0003-0395-8372
University of Cambridge, Cambridge, United Kingdom

Afra Pujol i Campeny  orcid.org/0000-0003-2895-2989
University of Cambridge, Cambridge, United Kingdom; Université Côte d'Azur, UMR 7320 – Bases, Corpus, Langage, Nice, France; University of Alicante, Alicante, Spain

- Batllore, M., Iglésias, N., & Martins, A. M.** (2005). Sintaxi dels clitics pronominals en català medieval. *Caplletra. Revista Internacional de Filologia*, 38, 137–177. URL: <https://ojs.uv.es/index.php/caplletra/article/view/4877>, number: 38. DOI: <https://doi.org/10.7203/caplletra.38.4877>
- Bieber, D., & Conrad, S.** (2009). *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511814358>
- Bruguera, J.** (1991). *El Llibre dels Fets del Rei en Jaume*. Barcelona: Barcino.
- Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., & Panchenko, A.** (2019). Targer: Neural argument mining at your fingertips. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 195–200). DOI: <https://doi.org/10.18653/v1/P19-3031>
- Clérice, T., Pilla, J., FrFerry, Camps, J.-B., ngawangtrinley, architexte, Jetely, A., Pinche, A., & Siddhant, S.** (2021). hipster-philology/pyrrha: 3.0.0. DOI: <https://doi.org/10.5281/zenodo.5144781>
- Colomina i Castanyer, J.** (2008). Paradigmes flectius de les altres classes nominals. In *Gramàtica del català contemporani, III*, 570–579. Barcelona: Editorial Empúries.
- Colón Domènech, G.** (Ed.) (2012). *El Llibre dels Feits: Aproximació Crítica*. No. 10 in Col·lecció Actes. València: Acadèmia Valenciana de la Llengua.
- Daelemans, W., Zavrel, J., Van den Bosch, A., & Van Der Sloot, K.** (2010). Mbt: memory-based tagger. *Version*, 3, 10–04.
- Fischer, S.** (2011). *The Catalan Clitic System*. Mouton de Gruyter. URL: <https://www.degruyter.com/document/doi/10.1515/9783110892505/html>, publication Title: The Catalan Clitic System.
- Institut d'Estudis Catalans, I. d. C.** (2016). *Gramàtica de la llengua catalana*. Barcelona: Institut d'Estudis Catalans. Publication Title: Gramàtica de la llengua catalana.
- Koch, P., & Oesterreicher, W.** (2012). Language of Immediacy – Language of Distance: Orality and Literacy from the Perspective of Language Theory and Linguistic History. In C. Lange, B. Weber, & G. Wolf (Eds.), *Communicative Spaces: Variation, Contact, and Change- Papers in Honour of Ursula Schaefer* (pp. 441–473). Bern: Peter Lang. URL: <https://www.peterlang.com/view/9783653021783/9783653021783.00028.xml>, publication Title: Communicative Spaces.
- Kytö, M., & Smitterberg, E.** (2015). Diachronic registers. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics, Cambridge Handbooks in Language and Linguistics* (pp. 330–345). Cambridge: Cambridge University Press. URL: <https://www.cambridge.org/core/books/cambridge-handbook-of-english-corpus-linguistics/diachronic-registers/B2B1B88606D7099492ECB1F591765347>. DOI: <https://doi.org/10.1017/CBO9781139764377.019>
- Meelen, M.** (2016). *Why Jesus and Job spoke bad Welsh: The origin and distribution of V2 orders in Middle Welsh*. LOT.
- Meelen, M., Roux, E., & Hill, N.** (2021). Optimisation of the largest annotated tibetan corpus combining rulebased, memory-based, and deep-learning methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1). DOI: <https://doi.org/10.1145/3409488>
- Partridge, S., & Kwakkel, E.** (2012). *Author, Reader, Book: Medieval Authorship in Theory and Practice*. Toronto: University of Toronto Press. URL: <https://www.degruyter.com/document/doi/10.3138/9781442665743/html>. DOI: <https://doi.org/10.3138/9781442665743>
- Pujol i Campeny, A.** (2018). *Word Order in Old Catalan*. Doctoral Thesis, University of Cambridge, Modern and Medieval Languages and Linguistics Faculty, Section of Theoretical and Applied Linguistics, Cambridge, United Kingdom. URL: <https://www.repository.cam.ac.uk/handle/1810/293049>
- Pujol i Campeny, A.** (2021). Code-switching in *Llibre dels Fets*: Language ideology in the 13th century Crown of Aragon. *Journal of Historical Sociolinguistics*, 7(1), 87–122. DOI: <https://doi.org/10.1515/jhsl-2019-0028>
- Rafel i Fontanals, J.** (2009). *Diccionari de Textos Catalans Antics*. URL: <http://www.ub.edu/diccionari-dtca/>
- Soldevila, F.** (1971). *Les quatre grans cròniques. Revisió del text, pròlegs i notes per Ferran Soldevila*. Barcelona: Selecta.
- Torruella, J., Pérez Saldanya, M., & Martines, J.** (2009). *Corpus Informatitzat del Català Antic*. URL: <http://cica.cat/>
- Vinas, A., & Vinas, R.** (2008). *El Llibre dels Fets de Jaume el Conqueridor*. Palma de Mallorca: Editorial Moll.
- Walkden, G.** (2016). The HeliPaD: A parsed corpus of Old Saxon. *International journal of corpus linguistics*, 21(4), 559–571. DOI: <https://doi.org/10.1075/ijcl.21.4.05wal>

TO CITE THIS ARTICLE:

Meelen, M., & Pujol i Campeny, A. (2021). Old Catalan Morphosyntax: Developing an Annotated Corpus. *Journal of Open Humanities Data*, 7: 30. pp. 1–12. DOI: <https://doi.org/10.5334/johd.54>

Published: 21 December 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.