# Comparing John Walker's 18th century grammatical categories against those of today with Treetagger

Francois **HUANG**
Blanche **MIRET**
Preethi **SRINIVASAN**
Dao **THAUVIN**
Université de Paris

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

## Introduction

As it is very well known, Treetagger (*Helmut Schmid*, 1996) is a tool that is widely used in the Natural Language Processing domain to annotate multilingual corpora with POS tags and lemma information. The purpose of this study is to find out whether Treetagger could learn the grammatical categories of John Walker's *Critical Pronouncing Dictionary* (1791). To achieve it, we will first train tree tagger with a tagset from Walker's Dictionary's grammatical categories, and then train a second one with a tagset based on the *Brown Corpus*, a text collection which contains 500 samples of tagged English-language text. Thus, we will compare the results of both tree tagger.

## Data

**TreeTagger with Walker's grammatical categories**

- Lexicon : Words defined by J. Walker in his dictionary

- Tagset : Walker's grammatical categories

**TreeTagger with Brown Corpus's grammatical categories**

- Lexicon : Words extracted from Brown Corpus

- Tagset : Grammatical categories from Brown Corpus

---

Training Data : Extract from Brown Corpus

Test Data : Two extracts: one from Brown Corpus, which is tagged and is different from the training data. The other one is from some definitions written by Walker in his dictionary

## Methods

- We created a **mapping table** from Brown Corpus tagset to Walker's tagset and **applied it on our training data**.

| Treetagger's POS tags | Brown Corpus's POS tags |
|---|---|
| art | at, dt |
| adj | jj, dti, ap, dts, jjt, jjr, jjs |
| pret | vbd, dod, bed, bedz, hvd |
| s | nn, nr, nns, nps |
| conj | in, cs, cc, dtx |
| prep | in, to |
| part.pass. | hvd, vbn, md |
| tp (third person) | vbz, bez, hvz, doz |
| v. | bem, vb, hv, be, do, ber |
| adv. | rb, ql, abx, abn, abl, ex |
| pron. | pps, ppo, pn, wps, ppl, wdt |
| part. | vbg |
| interj. | uh |

- **Trained TreeTagger** using our training data with Walker's tagset.

- **Trained TreeTagger** using our training data with Brown Corpus tagset .

- **Ran** both trained Taggers on our test data.

- **Compared** the Brown Corpus's test data **results** to the **original tagged corpus**.

- **Checked** the tags obtained by the other **result**.

## Remarks on Walker's Dictionary

- Brown Corpus's tagset and grammatical categories of Walker's dictionary **don't match** because many categories **changed** or simply **disappeared** over the course of time.

- 37895 words were tagged and 879 were not in Walker's Dicitonary.

- The tag **contraction** hasn't been used for all the words which required it.
Examples : *NE'ER* is a poetical contraction for Never. Walker declared it as an adverb.
*TA'EN* is a poetical contraction for Taken. Walker declared it as a contraction.

## Preparing TreeTagger

**About our lexicon from Walker's dictionary:**

- We **added** some proper nouns, punctuation marks and cardinal numbers from our training set.
- We **fused** some categories to match brown corpus categories
- Some categories have **disappeared**, such as solemn nominative plural ( Hapax: Ye ), and A negative (or privative termination) ( Hapax: Less ).
- 33688 words

**About our lexicon from Brown Corpus:**

- 56057 words

**About our training set :**

- about 1500 sentences
- 28639 words (including punctuation marks)

## Results

### Differences between Brown Corpus tagging and our tagging :

With Walker's tagset :

- 1590 failures on 2332 tags.
- Too much 's.' tag. 1300 over 1590 failures are about this tag.
- Only 8 over 19 tags have been used : adj., s., SENT, pron.pers., CD, NP, art., interj.

With Brown Corpus's tagset :

- 91 failures on 2332 tags.

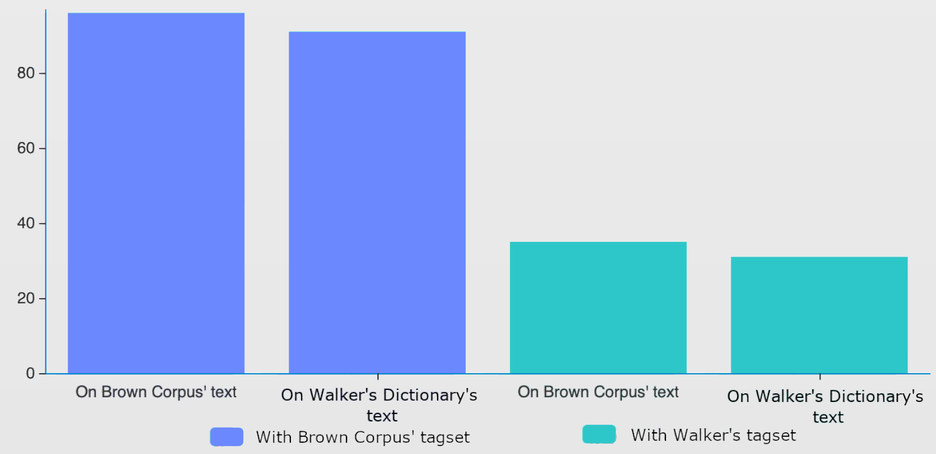
### Checking manually the tags obtained on Walker's Dictionary's definitions :

With Walker's tagset :

- 35 failures on 54 tags.
- 25 failures about the 's.' tag in our tagging.

With Brown Corpus's tagset :

- 5 failures on 54 tags.
- 2 failures about the ';' tag




**Comparaison of the percentage of correct tag between the two taggers on two corpus**

## Discussion about the results

For our Tagger with Walker's dictionary tagset :

- Size of the tagset : **19 tags** (while the Brown Corpus tagset has **86 tags**). Therefore, our **small number of tags** probably impacts the result.
- **Loss** of many useful tags from Walker's Dictionary. We could fix that by **tagging manually** the training set to keep most of the tags.
- Training on a **bigger training** set could give a better result. Most of the unused tags are rare tags that don't often appear in the training set. For example, the tag sp (for second person) appears only five times.

Finally, without taking into account those previous remarks, it seems that **Walker's dictionary tagset can't be applied to TreeTagger** with the same accuracy as it has with modern parameter files.

## References

1. J. Walker. *A Critical Pronouncing Dictionary*, 1824.

2. N. Trapateau. *A Critical Pronouncing Dictionary* from J. Walker xml file

3. W. N. Francis and H. Kucera. *Brown Corpus*

4. The project's Github : https://github.com/daothauvin/TreeTaggerWithWalker