Using TreeTagger trained with modern-language data on John Walker's 18th century dictionary

Francois **HUANG**
Blanche **MIRET**
Preethi **SRINIVASAN**
Dao **THAUVIN**
Univ Paris Diderot –
Sorbonne Paris Cité

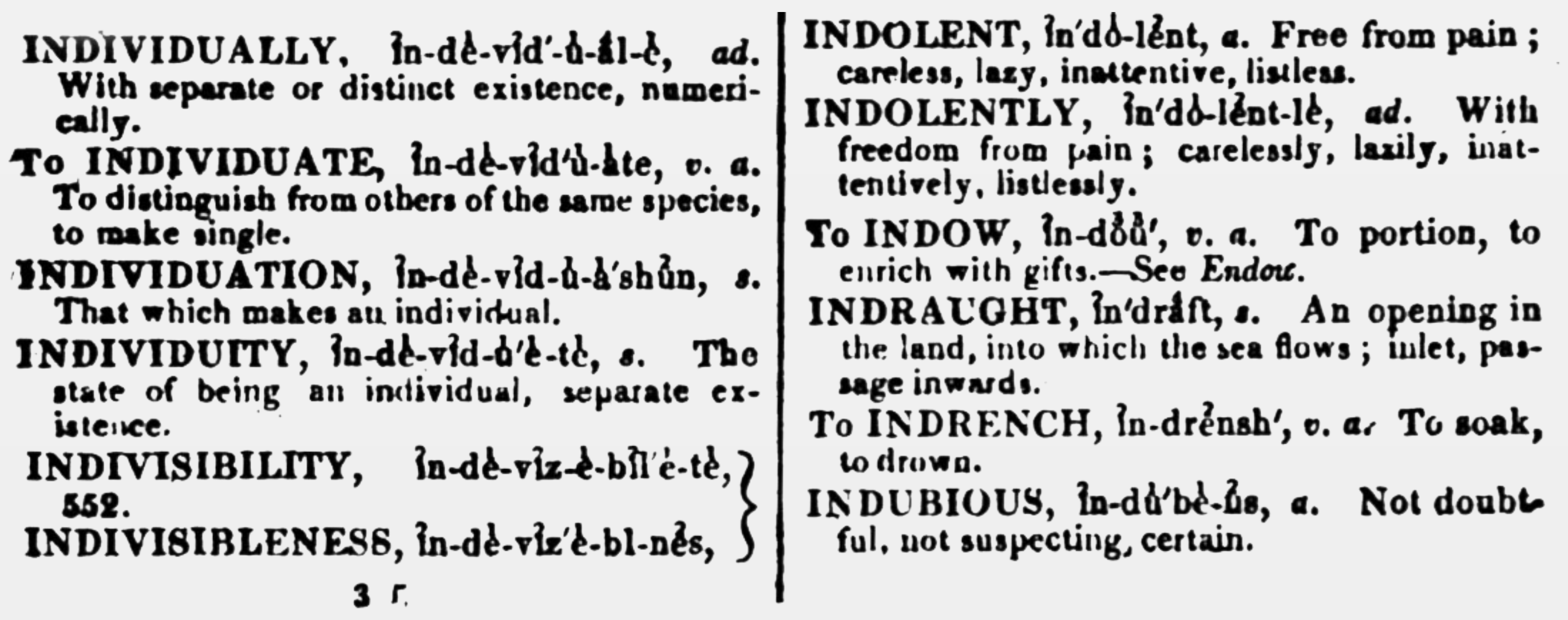
Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

Introduction

Since its development by Helmut Schmid in 1994, TreeTagger has successfully been used to tag various languages. The tool can be used to be trained with tagged data, then used to annotate a same language corpus with POS tags and lemma information. This research tends to analyse how TreeTagger, trained with modern-language data, reacts when executed on an 18th dictionary corpus. In particular, how would it respond when confronted with obsolete tokens? For instance we can find in the 18th century texts, tokens such as *advenes, affusing, gayety* which are no longer currently used; would it be able to correctly tag these tokens? More broadly, what information could the tool give about the evolution of a language over centuries?

Data and methods

The research is based on **John Walker's Critical Pronouncing Dictionary**(1), working on an XML file(2) representing the dictionary with machine-readable format.



INDIVIDUALLY, ìn-dè-vìd'-ù-âl-è, *ad.* With separate or distinct existence, numerically.
To INDIVIDUATE, ìn-dè-vìd'ù-âte, *v. a.* To distinguish from others of the same species, to make single.
INDIVIDUATION, ìn-dè-vìd-ù-à'shùn, *s.* That which makes an individual.
INDIVIDUITY, ìn-dè-vìd-ù'è-tè, *s.* The state of being an individual, separate existence.
INDIVISIBILITY, ìn-dè-vìz-è-bìl'è-tè, 552.
INDIVISIBLENESS, ìn-dè-vìz'è-bl-nès,
3 r.
INDOLENT, ìn'dò-lènt, *a.* Free from pain; careless, lazy, inattentive, listless.
INDOLENTLY, ìn'dò-lènt-lè, *ad.* With freedom from pain; carelessly, lazily, inattentively, listlessly.
To INDOW, ìn-dòù', *v. a.* To portion, to enrich with gifts.—See *Endow.*
INDRAUGHT, ìn'dråft, *s.* An opening in the land, into which the sea flows; inlet, passage inwards.
To INDRENCH, ìn-drènsh', *v. a.* To soak, to drown.
INDUBIOUS, ìn-dù'bè-ùs, *a.* Not doubtful, not suspecting, certain.

Extract from the *Critical Pronouncing Dictionary* (1824)

Number of words in data set:

386 172 words including **46 067 out of context words**, the ones defined by the dictionary and thus not presented in sentences. The rest are **in-context words**, included in phrases and therefore easier to tag for TreeTagger.

Procedure :

- **Adapt data** to TreeTagger-expected structure via some Python scripts

- **Run** TreeTagger

Hindsight on the result :

- The **precision** rate indicates the **accuracy** of the tagging calculated in the following way :

$$\frac{\#\text{ rightly tagged words}}{\#\text{ total words}}$$

- The **recall** rate indicates for each POS category how well it is **recognized**, again on a 200 words sample :

$$\frac{\#\text{ rightly identified words}}{\#\text{ words of the same POS category}}$$

Looking at words marked as **"unknown"** lemma is the lead to distinguish obsolete words.

Obsolete words :

A word is determined as obsolete in three ways, according to the *Oxford English Dictionary*:

- No dictionary entries found for the word

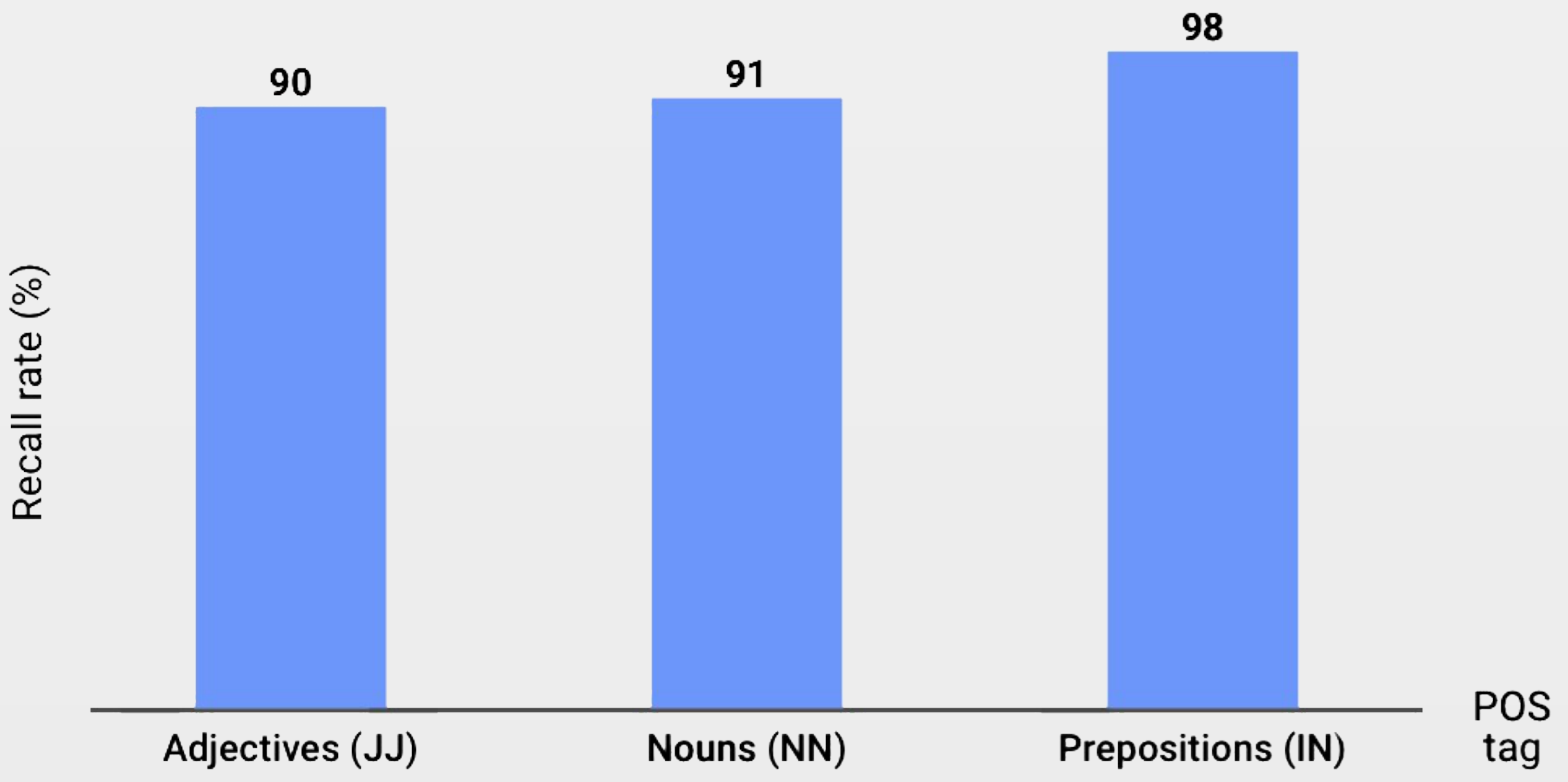
- Frequency (in current use) less or equal to $\frac{3}{8}$

- Marked as obsolete

Results

Word	Treetagger POS tag	Lemma
With	IN	with
separate	JJ	separate
or	CC	or
distinct	JJ	distinct
existence	NN	existence
,	,	,
numerically	RB	numerically
.	SENT	.

Precision rate of the result : **93,5%.**
Official rate of the TreeTagger on Penn-TreeBank data : **96,36%**(3)



Recall rate per POS category

Obsolete words :

Total <**unknown**> identified lemmas :

- **35 748** over 386 172 words

- Including **31 813 out-of-context words**

In a 200 **in-context** <**unknown**> token sample :

- **78,0%** are indeed obsoletes

- **66,7%** of these are correctly tagged

In a 200 **out-of-context** <**unknown**> token sample :

- **38,5%** are indeed **obsoletes**

- **0%** of these are **correctly tagged**

Word	Treetagger POS tag	Lemma
understeward	NN	unknown
calumniator	NN	unknown
unploughed	JJ	unknown
shortnecked	JJ	unknown
balsamick	NN	unknown
bathingroom	NN	unknown
absterging	VVG	unknown

Discussion about the results

Global precision is lower than official rate, which could signify some degree of language evolution.

What's interesting is the result about **obsolete words**. **9,2%** of the terms have been identified as **potentially obsolete**. However, the dataset being a **dictionary**, some of them were analyzed **out of context**. **89%** of the prospective **disused words** belong to this share of data and amongst these only **38,5%** are indeed **not used anymore**. But when regarding the **in-context words** exclusively, the success rate seems far more promising : **78%** are indeed **obsolete**. Furthermore, the **morpho-syntactic precision** of these obsolete words is **67%**, showing the **adaptation ability** of TreeTagger facing archaic vocabulary.

Remarks for extensive analysis :

1,2% of the in-context words were found potentially obsolete. To extend the research, one could bring some **new perspective** on the tested method in measuring the percentage of recognized obsolete words amongst a set certified as obsolete.

Conclusion

Ultimately, with an accuracy of 78% in using 'unknown' marked lemma as obsolete indicator, TreeTagger seems to be an interesting tool in measuring the evolution of the language, between two periods.

References

1. Walker J. (1824) *A Critical Pronouncing Dictionary & Expositor of The English Language*

3. Schmid, H. (1994). *Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing*, pages 44-49, Manchester, UK.

2. TRAPATEAU Nicolas (2015) *Placement de l'accent et voyelles inaccentuÃ©es dans la prononciation de l'anglais du XVIIIe siÃ¨cle sur la base du tÃ©moignage des dictionnaires de prononciation, des vers et de la musique vocale*, ThÃ¨se, universitÃ© de Poitiers.

4. D. Khurana, A. Koli, K. Khatter, S. Singh. *Natural Language Processing: State of The Art, Current Trends and Challenges*, Manav Rachma International University, 2017.

5. **The project's Github** : `https://github.com/BlancheMiret/TreeTagger_on_Walker`