Recovering from accidental syncretisms with word2vec analogies

Camille Lavigne, Gilles Boyé, Gauvain Schalchli

Recovering from accidental syncretisms

with word2vec analogies

Recovering from accidental syncretisms with word2vec analogies

Camille Lavigne, Gilles Boyé & Gauvain Schalchli

Université Bordeaux-Montaigne

CLLE-Montaigne

Systematic idiosyncratic syncretism

- It has received a lot of attention
  - Aronoff (1994) introduced morphomes to capture it
  - Maiden (1992) questioned its stability
  - Baerman et al. (2005) discuss its possible orientations
  - Bonami & Boyé (2003) studied its structure
- And with the emergence of the Paradigm Cell Filling Problem, systematic syncretism and the associated morphomes have become a central part of the discussion

The Paradigm Cell Filling Problem

(1) Some formulations:

a. *What licenses reliable inferences about inflected (and derived) surface forms of a lexical item?* (Ackerman et al. 2009)

b. *Given exposure to an inflected word form of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?* (Ackerman & Malouf 2013)

Some overlooked problems

- The Paradigm Cell **Finding** Problem (Boyé & Schalchli 2019):

  (2) *Given exposure to a sample of inflected forms with enough lexemes, what licences reliable inferences about the paradigm shape of these lexemes?*

- The scope of the Paradigm Cell **Filling** Problem:
  - should be based on the knowledge available to the speaker
  - should not be limited to novel lexemes

  (1b) *Given exposure to inflected wordforms of a lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?*

Accidental syncretisms

The Paradigm Cell Association Problem:

- *Assuming that speakers have identified a morphomic paradigm shape based on contrasting forms, how can they place (syncretic) forms in the morphomic paradigm?*

- Hypothesis:
  - High frequency lexemes are the key to finding the paradigm cells
  - they have the most contrasting forms => larger set of forms
  - they appear in the most varied contexts => well-defined distributions

➡ Speakers can use them to classify other forms and recover from the accidental syncretisms

A small experiment

English preterit & past participle

- A large proportion of English verbs have the same form for their preterit and their past participle
  - all regular verbs: walked, opened, enjoyed, …
  - many irregulars: thought, made, kept, …
- But among the most frequent verbs, it is common to have two different forms
  - ate-eaten, blew-blown, did-done, …
- We can use word-vectors to identify the morphomic cells occupied by a form in a corpus by comparing its word vector to our base of frequent contrastive items

Word2Vec

Relevance

- contextual embeddings
- no differenciation between homographic forms
- simple, efficient, available

Parameters

- skipgram negative sampling
- 152 dimensions
- 7-word window
- 20 tokens minimum
- 10 epochs
- negative sampling 15

Two corpora

- Brown
  - small corpus (1M words)
  - tagged
- BNC
  - large corpus (100M words)
  - divided in manageable parts
    - 100k to 15M words
  - tagged and untagged versions

Looking for a single shift vector

using the average shift from one set of verbs

Simple hypothesis for the tagged corpora: one shift vector fits all

1. we calculated the average $\vec{V}_{PPART}$ - $\vec{V}_{PRET}$ for a set of verbs ($\vec{V}_{SHIFT}$)
2. we added this average vector $\vec{V}_{SHIFT}$ to all the $\vec{V}_{PRET}$ available
3. cosine similarity results never converged uniformly
    - on the contrary, with larger sets of verbs, the cosine similarity was really bad

Looking for shift vectors

using multiple verbs as potential models

Second hypothesis for the tagged corpora: use a set of shift vectors

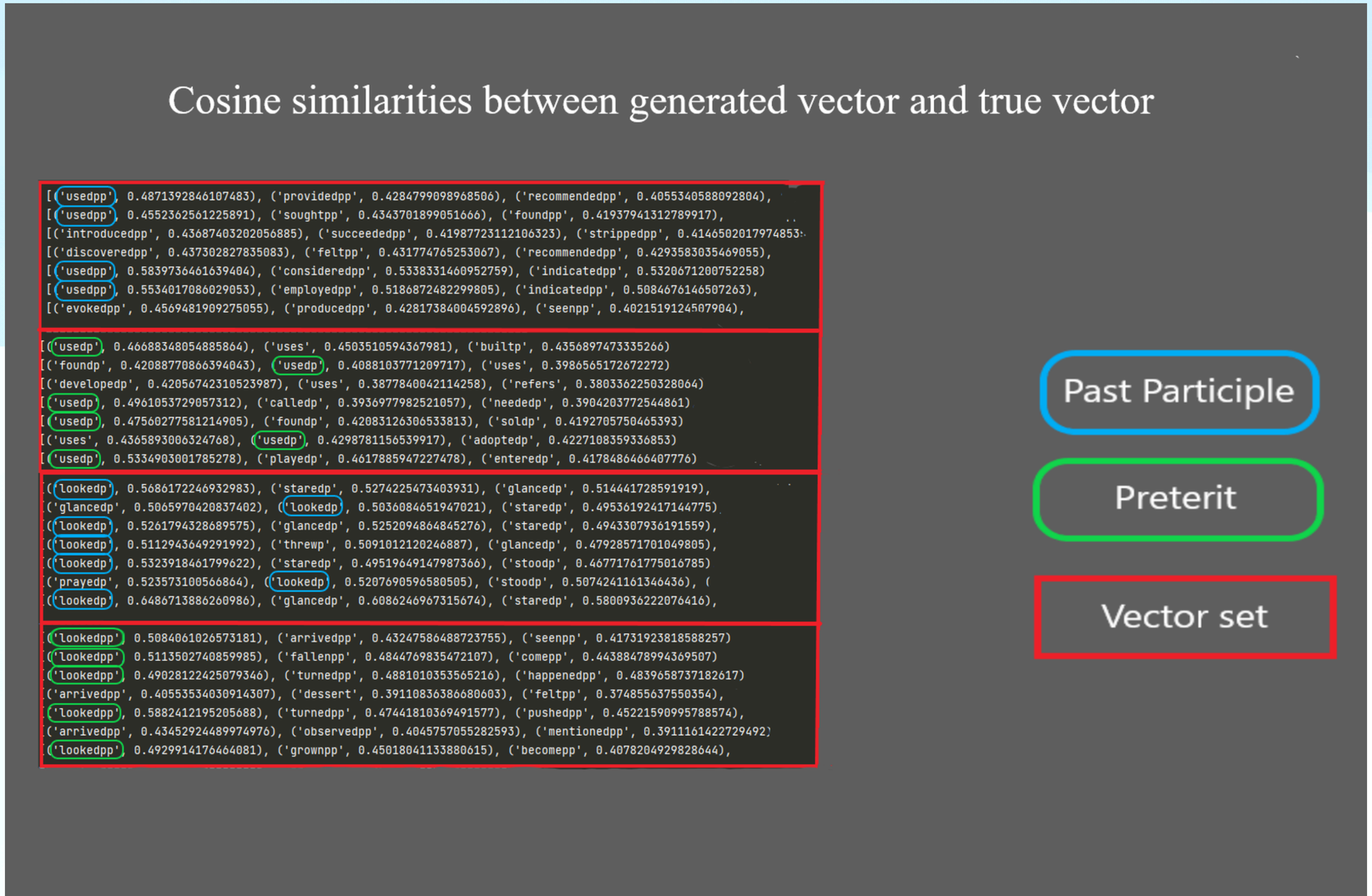
1. we calculated the $\vec{V}_{PPART}$ - $\vec{V}_{PRET}$ for a set of verbs ($\vec{V}_{SHIFT\_V}$)
2. we added all the shift vectors $\vec{V}_{SHIFT}$ to all the $\vec{V}_{PRET}$ available
3. every addition gives a candidate vector for the expected form

➡ in the resulting candidates, there were almost always one in close proximity to the target forms

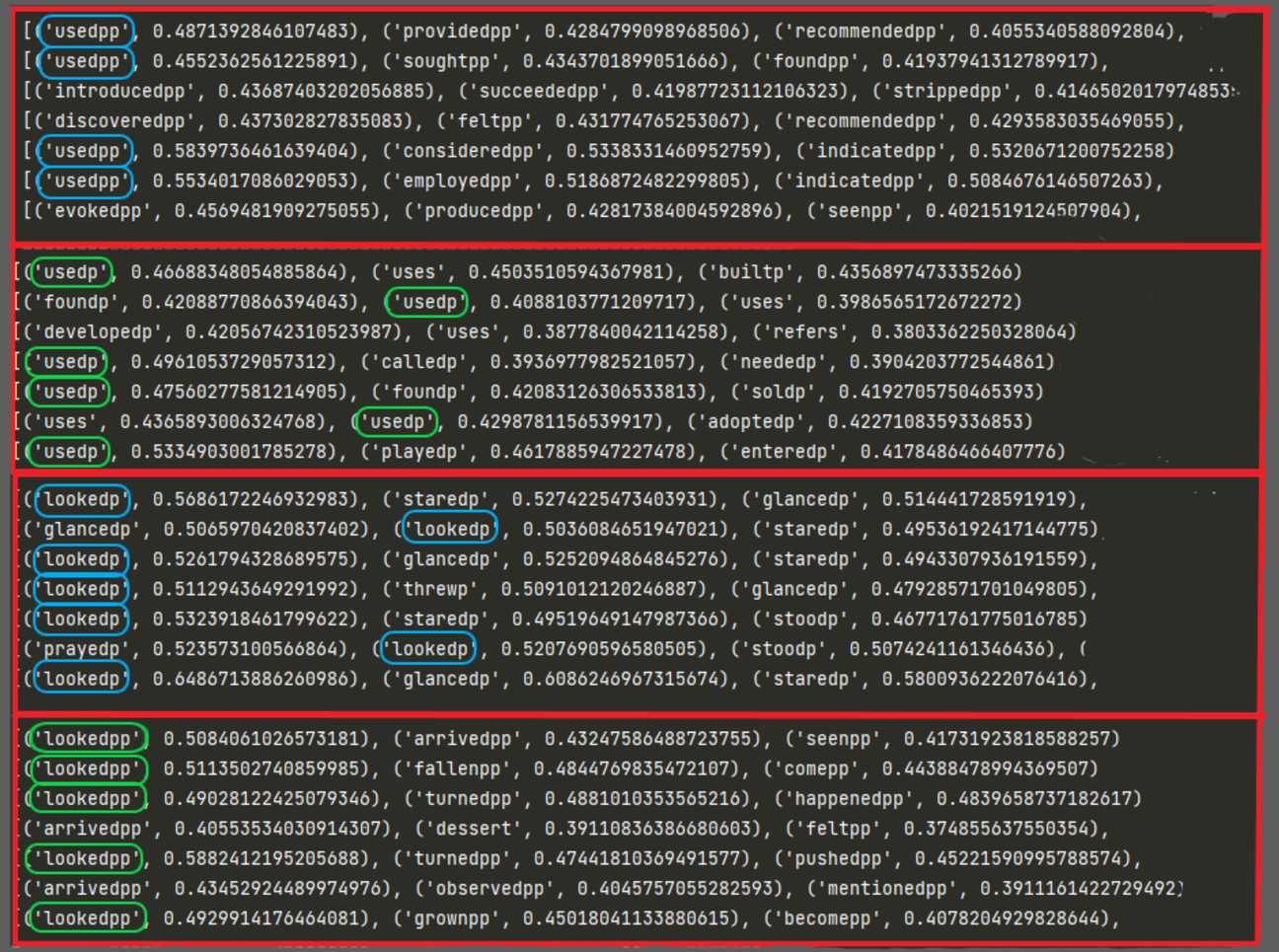
- not always the same using the same shift vector
- many times several shift vectors yielded good results

Shift vectors and their cosine similarities

using multiple verbs as potential models



Cosine similarities between generated vector and true vector



Past Participle

Preterit

Vector set

Predictions from the shift vector set

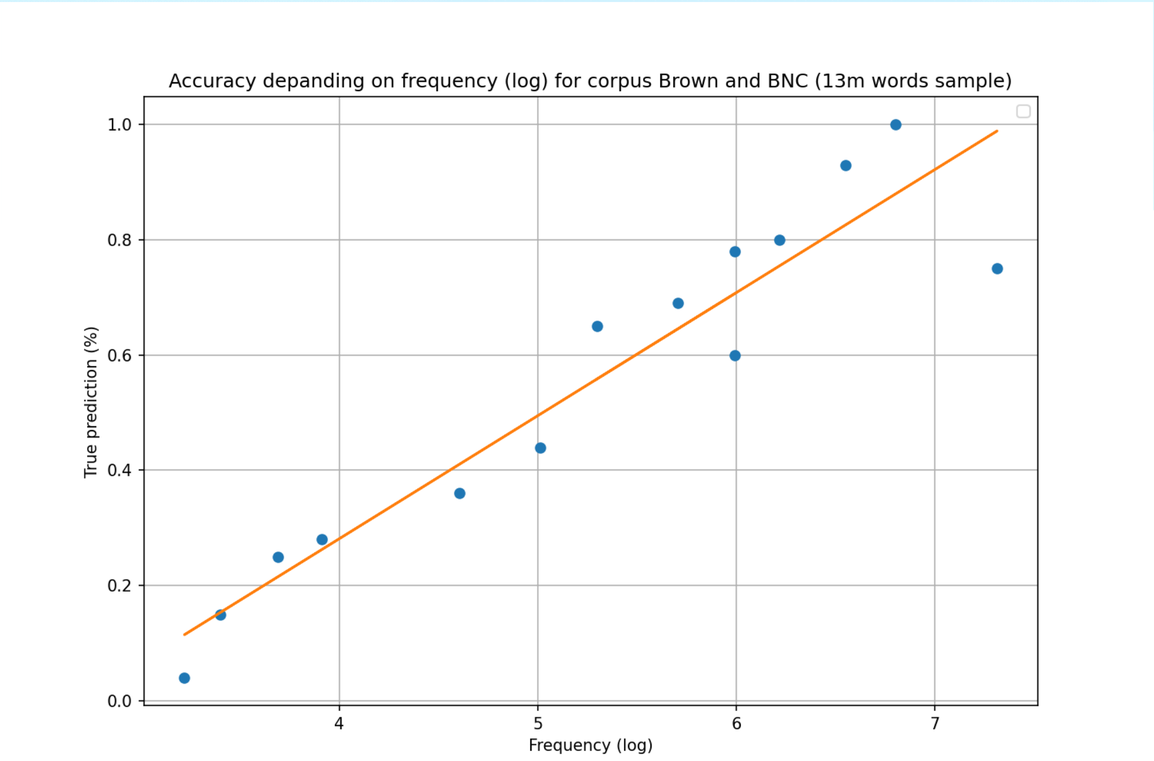
Problem: how to evaluate results?

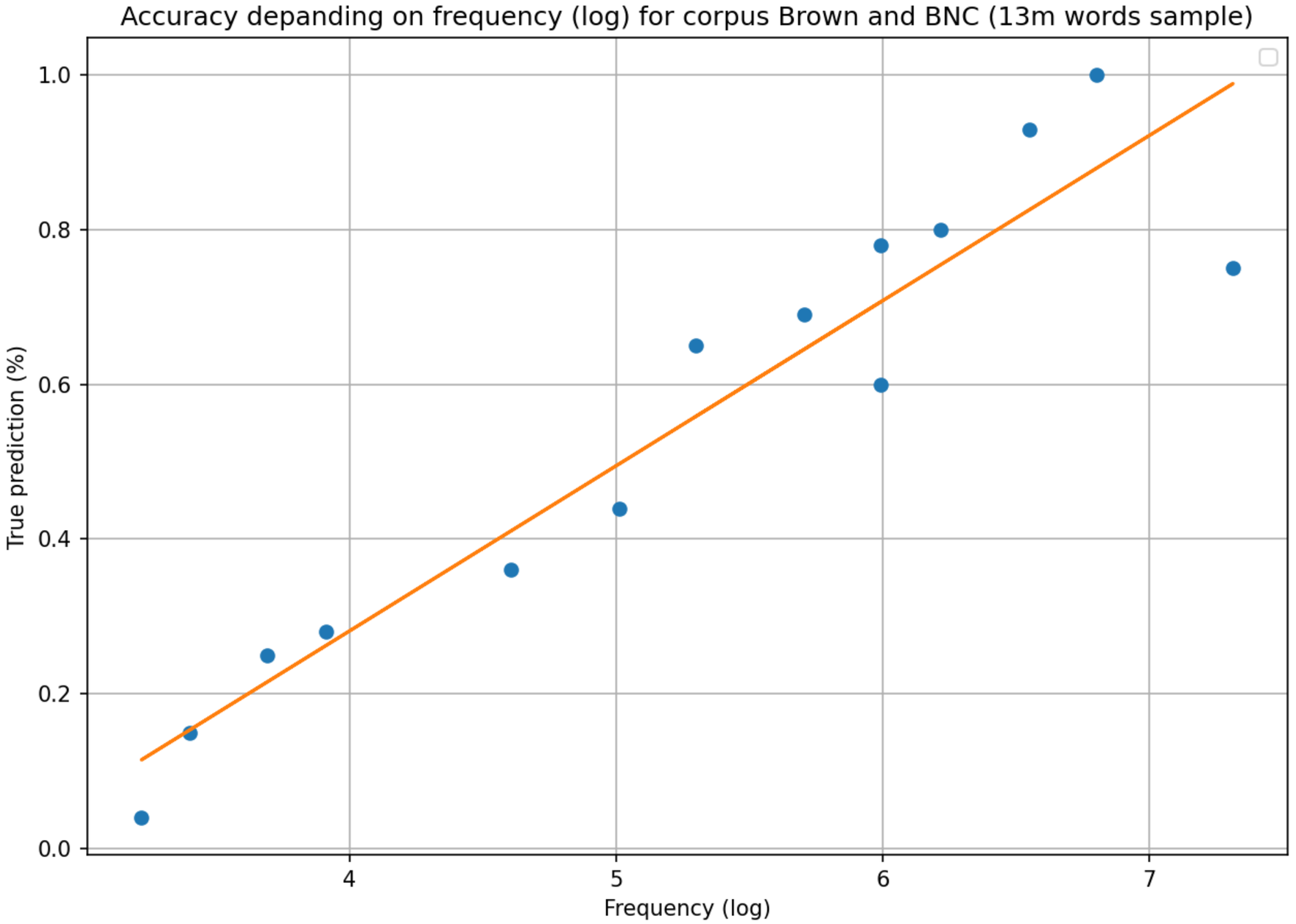
- Proposition

  1. take the k best cosine similarities of a given set
  2. count the corresponding forms of in the results
  3. count as true prediction when the most counted form is the expected one, otherwise count as fail

Results from the set of shift vectors

- accuracy is log-linearly correlated with the frequency of a given verb (r=0.94, p-value<10^{-6})
- overall accuracy is 25% only on Brown (probably too small) but reaches 81% on the BNC





LinregressResult(slope=0.21350889635596002, intercept=-0.5725907469078314, rvalue=0.945112792670326, pvalue=3.504946844732338e-07, stderr=0.021308342717363406, intercept_stderr=0

Summing up the exploration of the tagged corpora

- No single shift vector solution seems viable but finding seperate shift vectors for different groups of syncretic forms seems possible
    - but some clustering methods should help determine which shift vector is relevant for which syncretic form

Next hypothesis for the **untagged** corpus using the **tagged** corpus for control:

- *if a syncretic form is most similar to the sum of the vectors of the preterit and the past participle of a discriminating verb in the **untagged corpus**, the shift vector between the discriminating forms in the **tagged corpus** applied to the tagged preterit should be in close proximity to the tagged past participle*

The vectors in the untagged corpus

- a base of 10 discriminating verbs with:
  - 1 vector $\vec{V}_{R1}$ for the preterit (e.g. ate)
  - 1 vector $\vec{V}_{R2}$ for the past participle (e.g. eaten)
  - 1 vector $\vec{V}_{R3}$ for the sum of the two $\vec{V}_{R1} + \vec{V}_{R2}$ (e.g. ate+eaten)

  ➡ 30 reference vectors
- 190 syncretic verbs
  - 1 vector $\vec{V}_{S}$ for the common form of the preterit and the past participle

  ➡ 190 syncretic vectors

Finding the best analogy

- For each syncretic vector $\vec{V}_S$
  - we look for the closest reference vector $\vec{V}_R$
    - if it is a combination vector, $\vec{V}_{R3}$ ($\vec{V}_{R1}$ + $\vec{V}_{R2}$)
      - in the tagged corpus, we calculate the corresponding difference, $\vec{V}'_{S1}$ + $\vec{V}'_{R2}$ - $\vec{V}'_{R1}$, and look for the closest vectors in the tagged corpus
      - if the results are good enough ($\vec{V}'_{S1}$ + $\vec{V}'_{R2}$ - $\vec{V}'_{R1}$ ≅ $\vec{V}'_{S2}$), we can introduce pairs of separate vectors in the place of the original syncretic ones:
        - $\vec{V}_{S1}$ = ($\vec{V}_S$ + $\vec{V}_{R1}$ - $\vec{V}_{R2}$)/2
        - $\vec{V}_{S2}$ = ($\vec{V}_S$ + $\vec{V}_{R2}$ - $\vec{V}_{R1}$)/2

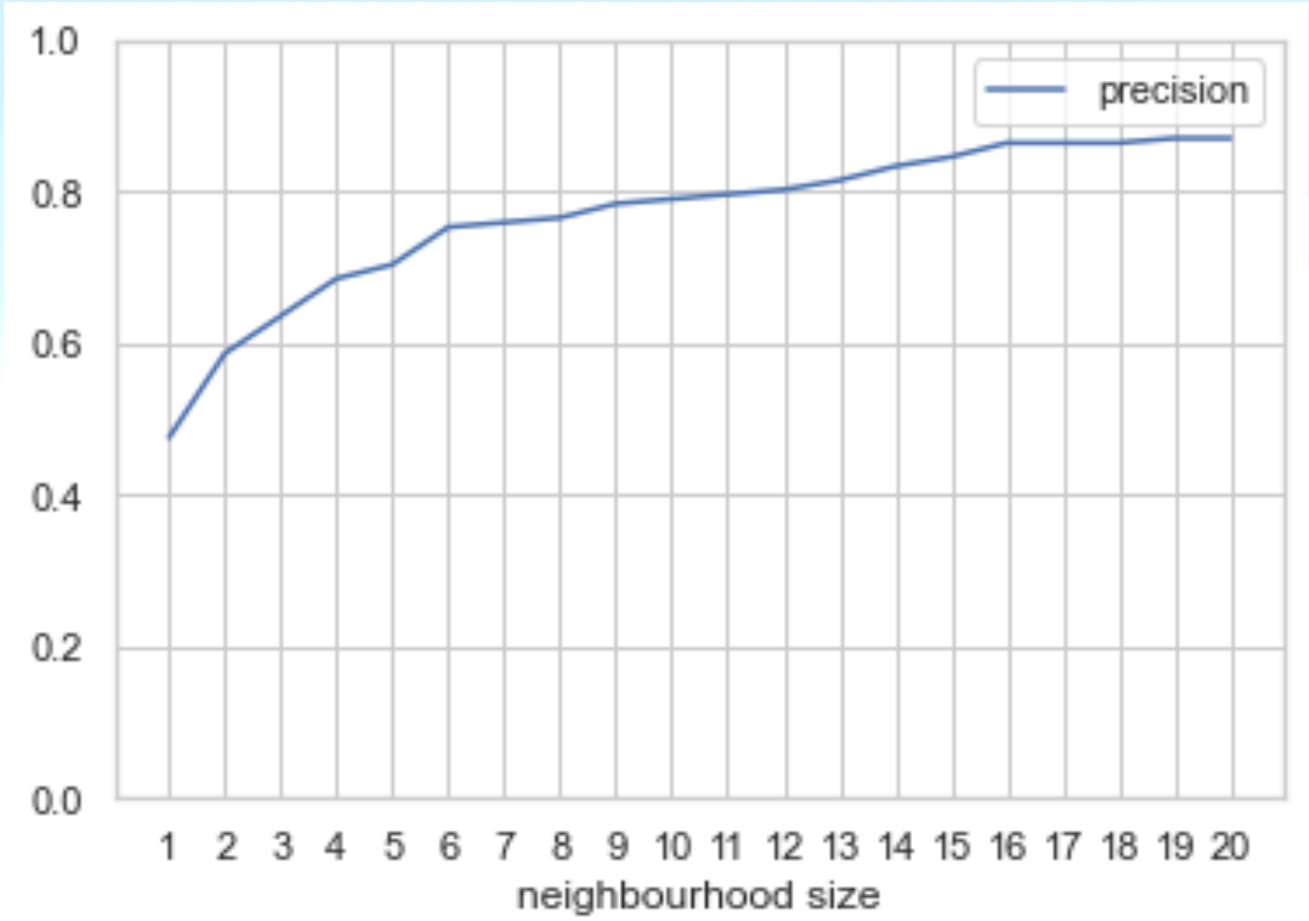
Classification results

reference vectors	syncretic vectors	proximity rank	nb
ate+eaten	found, used, provided, included, produced, bou...	51, 242, 682, 841, 287, 8, 156, 79, 98, 248, 2...	26
blew	opened	38	1
blew+blown	shot	2	1
did	wanted, let, refused	10, 68, 90	3
did+done	needed, helped, tried, meant, understood, paid...	60, 108, 20, 6, 65, 235, 53, 45, 191, 7, 282, ...	14
drew+drawn	made, brought, set, held, obtained, sought, in...	32, 1, 11, 66, 222, 229, 607, 104, 4, 88, 406,...	21
drove	started, stopped, stayed, returned, watched, hit	42, 15, 55, 35, 11, 7	6
drove+driven	moved, followed, offered, called, built, creat...	7, 16, 436, 24, 30, 529, 99, 616, 23, 378, 10,...	21
forgot	had, said, felt, talked	18, 43, 11, 2	4
forgot+forgotten	thought, told, liked, believed, remembered, he...	6, 25, 12, 20, 7, 0, 330, 56, 116, 271, 19, 31...	40
grew+grown	developed, remained, increased, tended, result...	353, 64, 775, 124, 130, 36	6
sank+sunk	formed, reached, settled, dropped, fought, dis...	176, 18, 22, 1, 75, 8, 6	7
spoke	seemed, lived, appeared, met, played, joined, ...	14, 36, 14, 20, 155, 47, 23, 218, 223, 55, 106	11
spoke+spoken	asked, read, described, shared, referred, cont...	48, 11, 731, 260, 145, 568, 267, 12, 23, 405, ...	14
threw	looked, stood	16, 14	2

Precision and neighbourhood size

Depending on the neighbourhood size chosen the evaluation of the precision varies:

- in 50% of cases, the closest vector to the computed one is the expected one
- in 90% of cases, the expected vector is in the 20 closest neighbours



Conclusion

- no single shift vector fitted all our syncretic verbs
  - but using a base of reference vectors and finding the closest reference allowed us to find sets of syncretic forms sharing the same shifting vector and recover from the accidental syncretisms
- This is exploratory work
  - the dataset was small, the syncretism was simple and we did not look at the error distribution
  - we would like to try this on French accidental syncretisms in conjugation where the patterns are more complex and the references are more ambiguous

Thank you