Striking out on one's own

Idiosyncratic frequency as a measure of derivation vs inflection

Maria Copot, Timothee Mickus & Olivier Bonami

Outline

Outline

Inflection vs Derivation

- Derivation increasingly recognised as paradigmatic, in a parallel way to inflection
    - See among many others: Marle (1984), Becker (1993), Bochner (1993), Blevins (2001), Stump (2005), Stekauer (2014), Boyé and Schalchli (2016), and Bonami and Strnadová (2019)
- A movement towards a unified, gradient approach based on empirical evidence.

The theory

- While not a dichotomy, **inflection and derivation remain two distinct concepts in theory**

Inflection	Derivation
lire∼lisait	*lire∼lisible*
Outputs realisations of a single lexeme	Outputs independent lexical entries
Same concept	Different concepts

- Can this theoretical difference manifest itself empirically?

A general prediction

- Derivational output is inherently more independent from its base. More variability for members of derivational relationships.
- For example, meaning relationships are more predictable in inflection than derivation (Bonami and Paperno, 2018)

A more specific prediction

- For related reasons, we can expect a **difference in the predictability of word frequency** for inflection and derivation.
- Because **derived lexemes are independent lexical entities**, we expect their frequency to vary independently of their base

Verb	Action noun	Freq. ratio
ouvrer ‘to work’	*ouvrage* ‘work; book’	0.02
cambrioler ‘to rob’	*cambriolage* ‘robbery’	0.34
	median	**17**
arriver ‘to arrive’	*arrivage* ‘delivery’	489
fixer ‘to fasten’	*fixage* ‘fixing’	1927

- In inflection, we do not expect such variability, except where it is semantically motivated (e.g. *eye* is more likely to be found in the plural than *nose*)

The research question

- Is the frequency of the output more predictable for inflection, compared to derivation?
    - Gradient vs dichotomy?
    - What factors are most helpful?

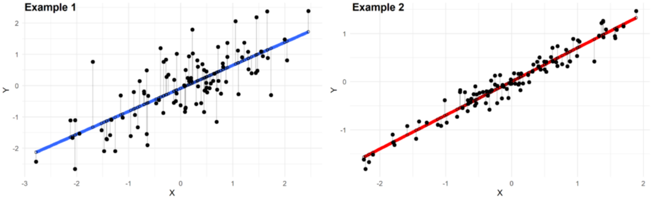
Outline

The methodological plan

How does a given morphological process impact frequency?

- We can train a statistical model for each morphological process to predict the frequency of the output

- We can use goodness of fit measures to compare our different models, and highlight whether some processes are harder to model than others
    - The residual standard error (RSE) of a model quantifies the accuracy of the prediction (low RSE = good prediction)



- …But what predictors should be used?

Frequency

We can use the frequency of a related form for a rough estimate of how frequently the lexeme is used

- We'll use the reference form (verb inf. & noun sg.)

reference form = citation form (of the base)

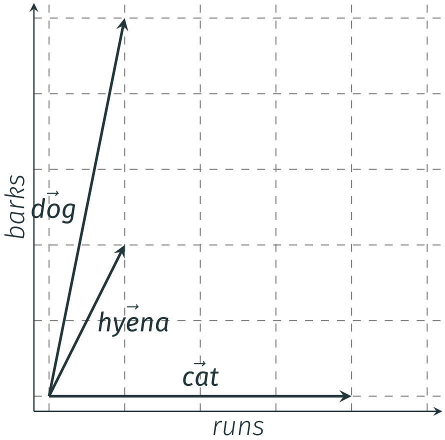
To compute frequencies, we need a large corpus:

- FrCOW 16: 6B tokens, in French, crawled from the web.
- We use the tokenization provided in the XML files.

“Word vectors”

- We want predictors for semantic information

	runs	*barks*
dog	1	5
hyena	1	2
cat	4	0

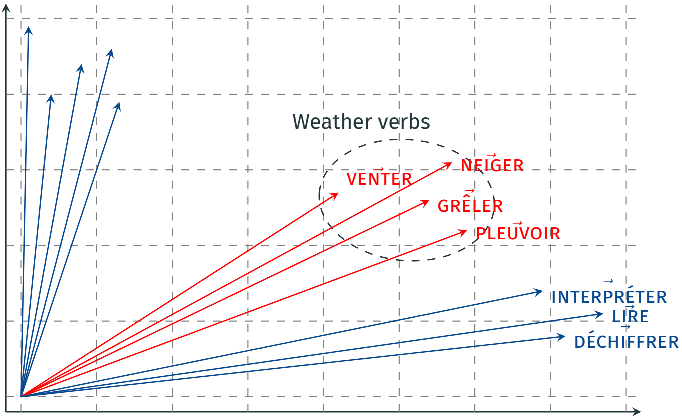


from Baroni, Bernardi, and Zamparelli (2014)

- We can use many different observations, e.g., on words or lexemes
- Frequency is encoded in the length (norm) of a vector

Semantic neighbourhood

- We expect the neighbours of a given word to **share semantic characteristics** with it



Word vectors: to recap

- word vectors reflect lexical semantics
- Regions of the semantic space describe coherent semantic fields (e.g., weather verb vectors are bunched together).

We can use vectors to make **semantically informed predictions.**

- We can use them **directly**: plug in the vector $\vec{w}$ of the word w
  (**Word-level** semantic information)
- We can use them **indirectly**: explore the neighbourhood of $\vec{w}$ which describe the general trend for semantically similar words
  (**Lexeme-level** semantic information)

We'll train **two 100D vector spaces** on FrCOW16 data.

Two many vector components

We want to use vectors as predictors in statistical models

> *"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."*
>
> John von Neumann

- If we use all 100 vector components, we would have more predictors than distinct responses.
- We can apply dimensionality reduction to solve this issue.

Dimensionality reduction with SVD

Click here for an amazing GIF that we couldn’t be bothered to embed properly!

Outline

The models

- A baseline: $f(output) \sim f(\text{reference form})$

The models

- A baseline: $f(output) \sim f(\text{reference form})$
  - $f(lirai) \sim f(lire)$

The models

- A baseline: $f(output) \sim f(\text{reference form})$
- $f(output) \sim f(\text{reference form}) + \overrightarrow{reference}\, form$

The models

- A baseline: $f(output) \sim f$(reference form)
- $f(output) \sim f$(reference form) + $\overrightarrow{reference\ form}$
  - $f(lirai) \sim f(lire) + \overrightarrow{\text{LIRE}}$
  - why? A way to take base semantics into account
    - Necessary to account for *eye*∼*eyes*, *nose*∼*noses*

The models

- A baseline: $f(output) \sim f(\text{reference form})$
- $f(output) \sim f(\text{reference form}) + \overrightarrow{reference\ form}$
- $f(output) \sim f(\text{reference form}) + \text{average neighbour relative frequency}$

The models

- A baseline: $f(output) \sim f($reference form$)$
- $f(output) \sim f($reference form$)$ + $\overrightarrow{reference\ form}$
- $f(output) \sim f($reference form$)$ + average neighbour relative frequency
  - average neighbour relative frequency = $\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{neighb\ form_i}{neighb\ ref\ form_i}$
  - $f(lirai) \sim f(lire) + avg(\frac{f(intérpreterai)}{f(intérpreter)} + \frac{f(déchiffrerai)}{f(déchiffrer)} + ...)$
  - why? For processes whose output is heavily dependent on the base, this should provide an accuracy boost.

The models

- A baseline: $f(output) \sim f(\text{reference form})$
- $f(output) \sim f(\text{reference form}) + \vec{reference\ form}$
- $f(output) \sim f(\text{reference form}) + \text{average neighbour relative frequency}$
- $f(output) \sim f(\text{reference form}) + \text{average } \vec{neighbour}$

The models

- A baseline: $f(output) \sim f(\text{reference form})$
- $f(output) \sim f(\text{reference form}) + \vec{reference\ form}$
- $f(output) \sim f(\text{reference form}) + \text{average neighbour relative frequency}$
- $f(output) \sim f(\text{reference form}) + \text{average } \vec{neighbour}$
    - $f(lirai) \sim f(lire) + \text{avg}(\vec{intérpreterai} + \vec{déchiffrerai} + \ldots)$
    - Neighbours of the base are obtained. The vector of their output is averaged and added as a predictor.
    - why? Same reason for model type 3, but semantics is included more directly.

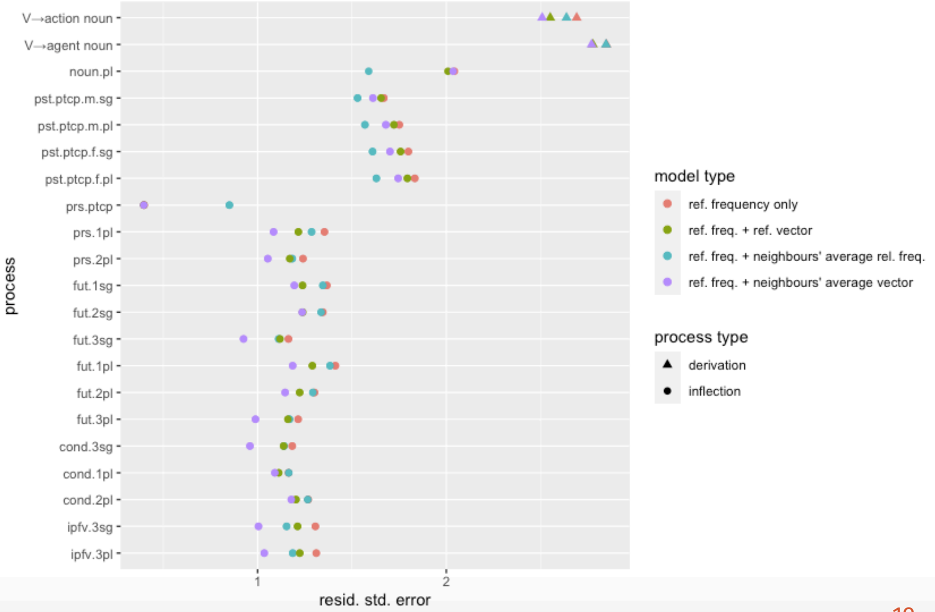
The data

- What morphological processes did we look at?
- Derivation[1]:
  - V $\longrightarrow$ ACTION NOUN
  - V $\longrightarrow$ AGENT NOUN
- Inflection[2]
  - Noun pluralisation
  - 18 verbal inflectional cells (excluded cells with high intraparadigmatic homophony, as frequency counts are unreliable)

[1] Datasets of derivational pairs are scarce, so we were not able to include more. Derivational pairs were selected from Demonette (Hathout and Namer, 2014)

[2] Inflectional pairs were based on the GLàFF (Sajous, Hathout, and Calderone, 2014)

Crunching the numbers



Conclusion

- Theoretically, the distinction between inflection and derivation is quite clear:
  - **Inflection**: different ways to talk about the same concept depending on context
  - **Derivation**: different concepts
- Prediction: qualities of derivational output are harder to predict from the base, compared to inflection. **This is borne out**: all of inflection has a lower RSE than all of derivation.
- The method employed shows promise for **better understanding the nature of different processes**.
  - For past participles, the output has inherently varied semantics, which is why models based on frequency rather than vectors are better predictors

Outline

Data selection

- Which cells in the paradigm of French verbs can we work with?
- Working with our dataset, we exclude…

Finite forms

	1SG	2SG	3SG	1PL	2PL	3PL
IND.PRS	2	3	183	2	5	14
IND.IPFV	0	0	5083	10	10	5076
IND.PST	4484	4448	4694	5116	5116	5101
FUT	5211	5207	5213	5190	5212	5221
SBJV.PRS	0	250	2	8	7	13
SBJV.IPFV	4701	4725	5119	4726	4738	4740
COND	0	0	5220	5212	5212	5215
IMP	—	0	—	2	2	—

Nonfinite forms

INF	PRS.PTCP	PST.PTCP			
		M.SG	F.SG	M.PL	F.PL
5006	4311	3935	3055	2903	3199

Number of verbs from Flexique with no homograph documented in the GLÀFF, by paradigm cell

Data selection

- Which cells in the paradigm of French verbs can we work with?
- Working with our dataset, we exclude...
  - cells with high numbers of homographs according to the GLÀFF;

Finite forms

	1SG	2SG	3SG	1PL	2PL	3PL
IND.PRS	2	3	183	2	5	14
IND.IPFV	0	0	5083	10	10	5076
IND.PST	4484	4448	4694	5116	5116	5101
FUT	5211	5207	5213	5190	5212	5221
SBJV.PRS	0	250	2	8	7	13
SBJV.IPFV	4701	4725	5119	4726	4738	4740
COND	0	0	5220	5212	5212	5215
IMP	—	0	—	2	2	—

Nonfinite forms

INF	PRS.PTCP	PST.PTCP			
		M.SG	F.SG	M.PL	F.PL
5006	4311	3935	3055	2903	3199

Number of verbs from Flexique with no homograph documented in the GLÀFF, by paradigm cell

Data selection

- Which cells in the paradigm of French verbs can we work with?
- Working with our dataset, we exclude...
  - cells with high numbers of homographs according to the GLÀFF;
  - cells out of current usage (i.e. most attestations are likely to be archaic);

Finite forms

	1SG	2SG	3SG	1PL	2PL	3PL
IND.PRS	2	3	183	2	5	14
IND.IPFV	0	0	5083	10	10	5076
IND.PST	4484	4448	4694	5116	5116	5101
FUT	5211	5207	5213	5190	5212	5221
SBJV.PRS	0	250	2	8	7	13
SBJV.IPFV	4701	4725	5119	4726	4738	4740
COND	0	0	5220	5212	5212	5215
IMP	—	0	—	2	2	—

Nonfinite forms

INF	PRS.PTCP	PST.PTCP			
		M.SG	F.SG	M.PL	F.PL
5006	4311	3935	3055	2903	3199

Number of verbs from Flexique with no homograph documented in the GLÀFF, by paradigm cell

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)
    - *know–knows*, *dance—dances*:

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)
    - *know–knows*, *dance—dances*:
      knowing the meaning of the former entails knowing the meaning of the latter

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)
  - *know–knows*, *dance—dances*:
    knowing the meaning of the former entails knowing the meaning of the latter
  - *sell–seller*, but *dine–diner* or *see–seer*:

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)

  - *know–knows*, *dance—dances*:
    knowing the meaning of the former entails knowing the meaning of the latter
  - *sell–seller*, but *dine–diner* or *see–seer*:
    knowing the meaning of the former does **not** entail knowing the meaning of the latter

Bonami and Paperno (2018), I

- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)
    - *know–knows*, *dance—dances*:
      knowing the meaning of the former entails knowing the meaning of the latter
    - *sell–seller*, but *dine–diner* or *see–seer*:
      knowing the meaning of the former does **not** entail knowing the meaning of the latter
- Bonami and Paperno (2018) test whether this assumption is consistent with distributional semantics

Bonami and Paperno (2018), I

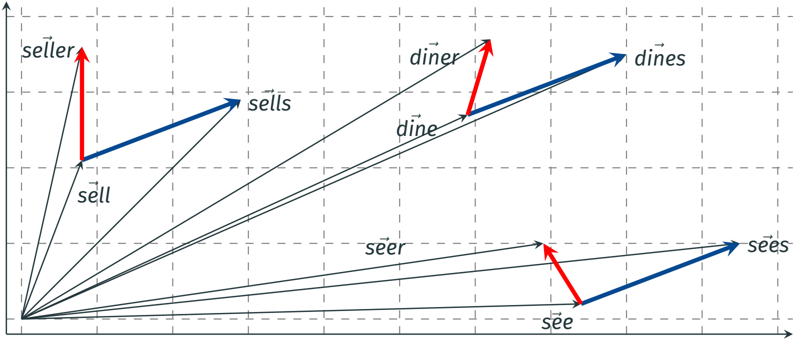
- In morphology, inflection has been claimed to be more semantically regular than derivation (Stump, 1998; Stekauer, 2014, e.g.)
  - *know–knows*, *dance—dances*:
    knowing the meaning of the former entails knowing the meaning of the latter
  - *sell–seller*, but *dine–diner* or *see–seer*:
    knowing the meaning of the former does **not** entail knowing the meaning of the latter
- Bonami and Paperno (2018) test whether this assumption is consistent with distributional semantics
- Assuming it is, we would expect linear offsets for inflectional relations (e.g., $\vec{bare} - \vec{3^{rd}sg}$) to be more consistent than those for derivational relations (e.g., $\vec{verb} - \vec{agent}$)

Bonami and Paperno (2018), II

- Many factors to control: frequency, but also the inherent semantics of the words under consideration

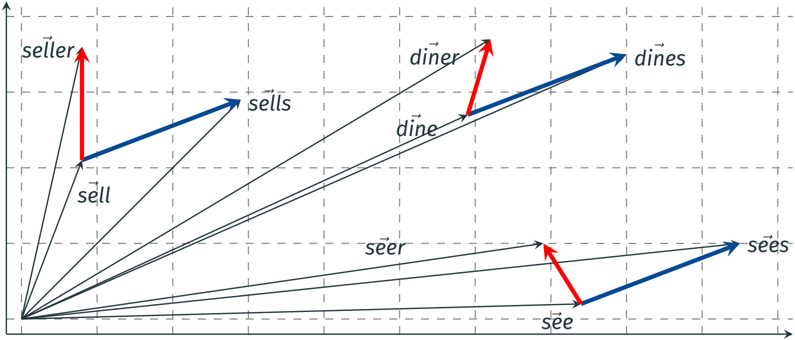
Bonami and Paperno (2018), II

- Many factors to control: frequency, but also the inherent semantics of the words under consideration
- The solution of Bonami and Paperno (2018) is to use word triples



Bonami and Paperno (2018), II

- Many factors to control: frequency, but also the inherent semantics of the words under consideration
- The solution of Bonami and Paperno (2018) is to use word triples



- They find that derivational relations yield significantly more variation than inflectional ones: derivational pairs stray more from the average value than inflectional pairs.