Concurrence et Polysémie
7 juin 2019 - LLF (UMR 7110)

Investigating rivalry through a distributional analysis

French –*age* suffixation vs. verb to noun conversion

Alice Missud & Florence Villoing
Université Paris Nanterre, MoDyCo (UMR, 7114)

DATA

Rival morphological schemas

- Derivation of nouns from verbs
- Use the same kind of semantic operations
- Example in French : the most extensively studied rivalry between *–age* and *–ment* suffixations because of large number of doublets.

(Tasmowski-de Ryck (1977), Dubois (1962), Kelling (2003), Martin (2007), Uth (2010, 2016), Fradin (2014, 2016), Dal et al. (2018).

(1) *battage / battement* 'beating' (of the heart), *décollage / décollement* 'unsticking', *emballage / emballement* 'wrapping',
gonflage 'pumping (of a tire) / *gonflement* 'expansion (of a tire)',
miaulage / miaulement 'mewing', *pavage / pavement* 'paving'

DATA

Rival morphological schemas

- Study of another productive rivalry neglected so far

→ French rivalry between **–*age* suffixation** and **verb to noun conversion**

HOW ARE THEY RIVALS?

Semantic interpretations

- Event (action) :
  (2) a. *-age* → *jardinage* 'gardening', *rasage* 'shaving', *ponçage* 'sanding'
  b. conversion → *survol* 'hovering', *baisse* 'drop', *secousse* 'shake'

- Object (result) :
  (3) a. *-age* → *bronzage* 'suntan', *maquillage* 'makeup', *héritage* 'inheritance'
  b. conversion → *empreinte* 'footprint', *certificat* 'certificate', *amas* 'heap'

HOW ARE THEY RIVALS?

Doublets

(4) *accrochage / accroche* 'hanging'

jetage / jet 'throwing'

pesage / pesée 'weighing'

rééquilibrage / rééquilibre 'rebalancing'

HOW ARE THEY RIVALS?

Doublets

(4) *accrochage / accroche* 'hanging'

jetage / jet 'throwing'

pesage / pesée 'weighing'

rééquilibrage / rééquilibre 'rebalancing'

Construction of nouns with the same semantic values

Selection of the same verbs

COMPETITION

STUDIES IN MORPHOLOGICAL RIVALRY

Rule-based representations

- Focusing on **input** and **output** constraints.

Selectional constraints on the verb

Semantic variation between derived lexemes

(Dubois 1962, Grimshaw 1990, Kelling 2003, Corbin 2004, Meinschäfer 2016, Martin 2007, Ferret et al. 2010)

- Problems: fail to account
  - for all the constraints that may apply
  - for the polysemy of deverbal nouns

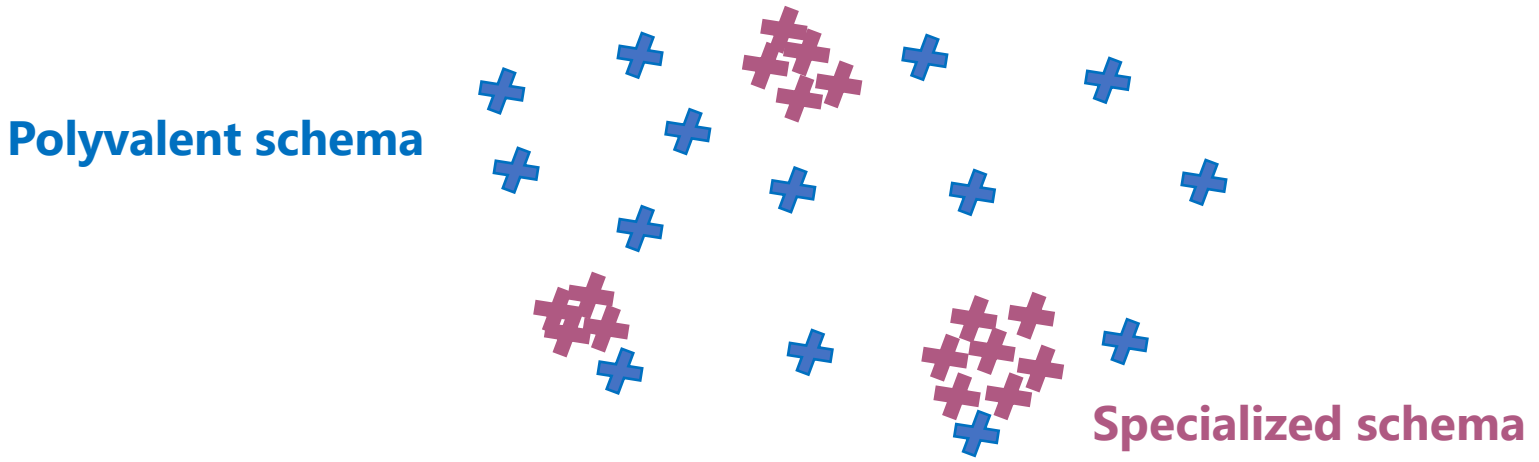
THEORETICAL FRAMEWORK

LEXEMATIC MORPHOLOGY

(Matthews 1974; Anderson 1992; Aronoff 1994; Fradin 2003; Booij 2005)

→ MORPHOLOGICAL NICHES

(Lindsay & Aronoff 2013; Arndt-Lappe 2014; Aronoff 2015, 2016, 2017, 2019)



THEORETICAL FRAMEWORK

QUANTITATIVE MORPHOLOGY
(Baayen 1992, 2009; Skousen 1989, 1992; Bonami & Thuilier 2019)

Probabilistic approach.
Looking for **statistical regularities** in the lexicon in order to find:

- Phonological constraints
- Morphological constraints
- Semantic constraints

HYPOTHESIS

The two schemas are distinguishable:

-*age* suffixation may be more constrained than conversion.

- Arguments -

Phonological constraints:

- Size constraint: -*age* might need shorter stems than conversion
- Dissimilation: no [ʒ] in the last syllable of the verb for –*age* (Lapraye 2017)

Morphological constraint

Semantic constraints

HYPOTHESIS

The two schemas are distinguishable:

-*age* suffixation may be more constrained than conversion.

- Arguments -

Phonological constraints

Morphological constraint:

- -*age* suffixation prefers verbs of the 1st group (Lapraye, 2017)

Semantic constraints

HYPOTHESIS

The two schemas are distinguishable:

-age suffixation may be more constrained than conversion.

- Arguments -

Phonological constraints

Morphological constraint

Semantic constraints:

- *-age* prototypically derives event or object nouns, conversion is more versatile → event, object, agent, instrument…

RESEARCH QUESTIONS

Focus on **semantic constraints** using **quantitative methods**.

→ How are *–age* derivatives and converted nouns distributed?

→ What are the semantic properties that drive this distribution?

Distributional properties (context → meaning):

- capture a lot of semantic information
- can be computed

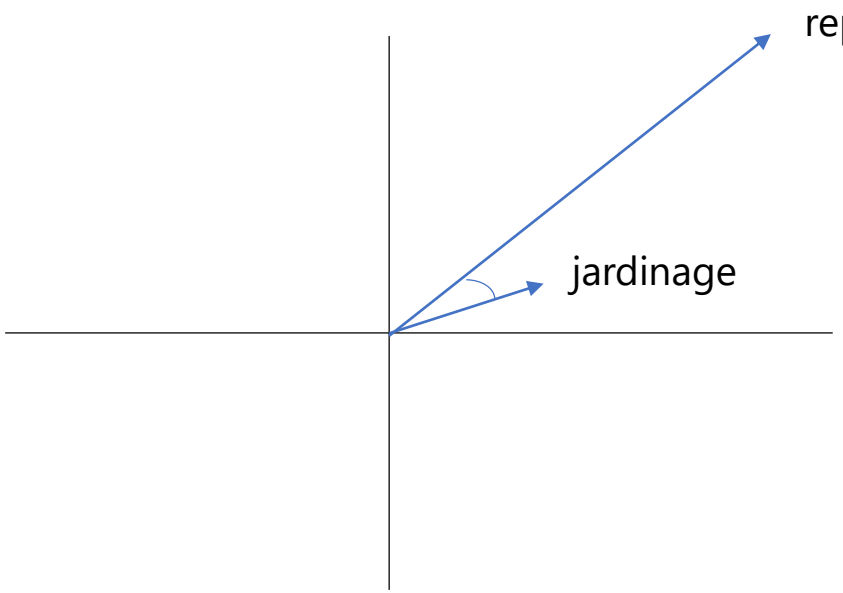
DISTRIBUTIONAL SEMANTICS

Word2Vec (Mikolov et al. 2013)

Neural network based model that reconstructs the meaning of a word using the words of its context.

Input: large corpora
Output: one word => one vector

Jardinage → [0.2, 0.3]
Repassage → [0.7, 0.8]



$$\cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}.$$

DATA

200 derived lexemes:

- 100 converted nouns from Tribout (2010)
- 100 –*age* suffixed nouns from FrWaC (1,6 billion words, Baroni et al. 2009)

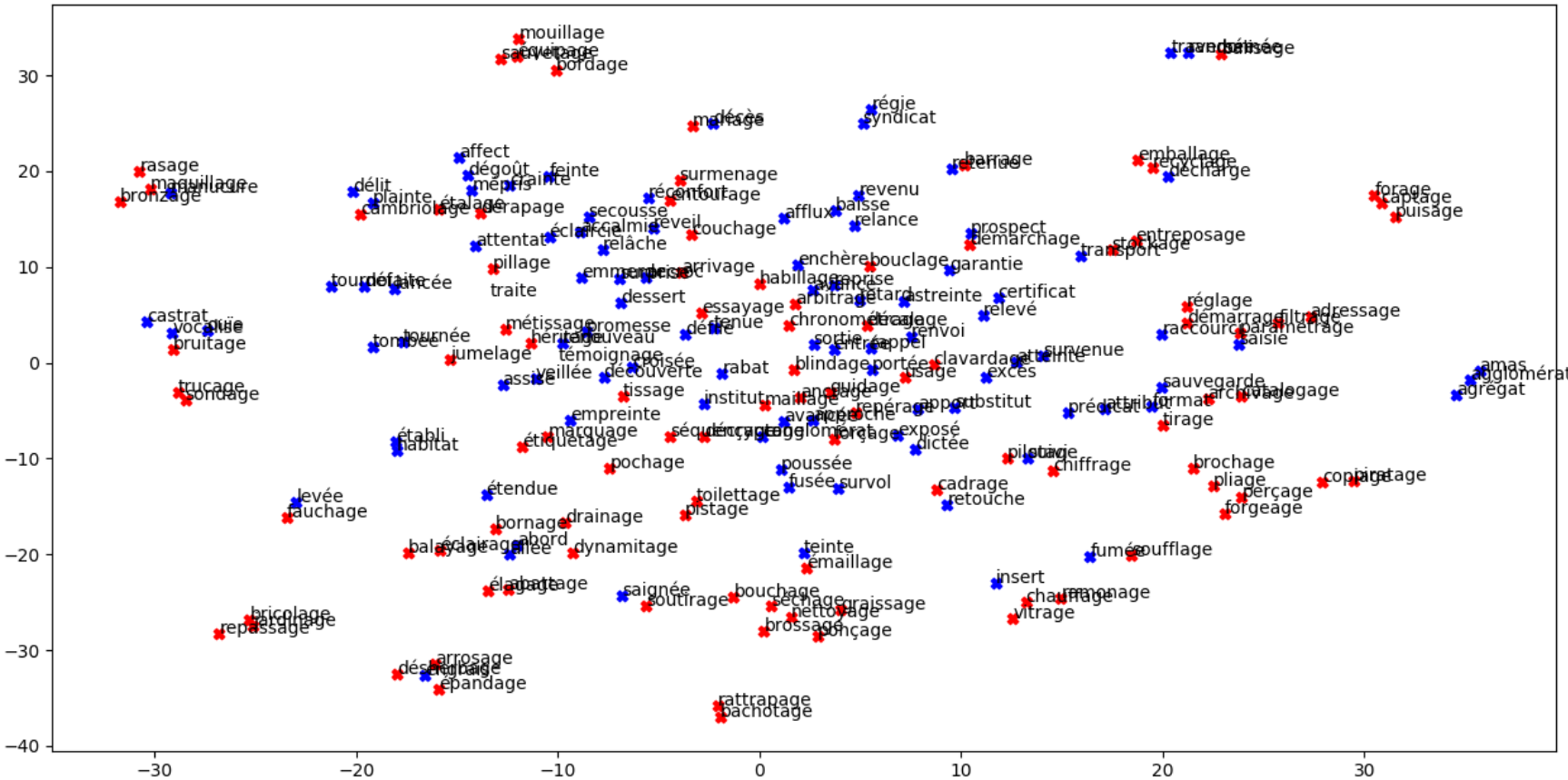
→ Most frequent nouns according to FrWaC frequencies
→ Excludes doublets like *rééquilibre* / *rééquilibrage*

Pre-trained vectors from Fauconnier (2015)

- Trained with Word2Vec using Skip-gram model
- On a POS tagged version of FrWaC
- Dimensions: 700

→ One vector for each derived lexeme.

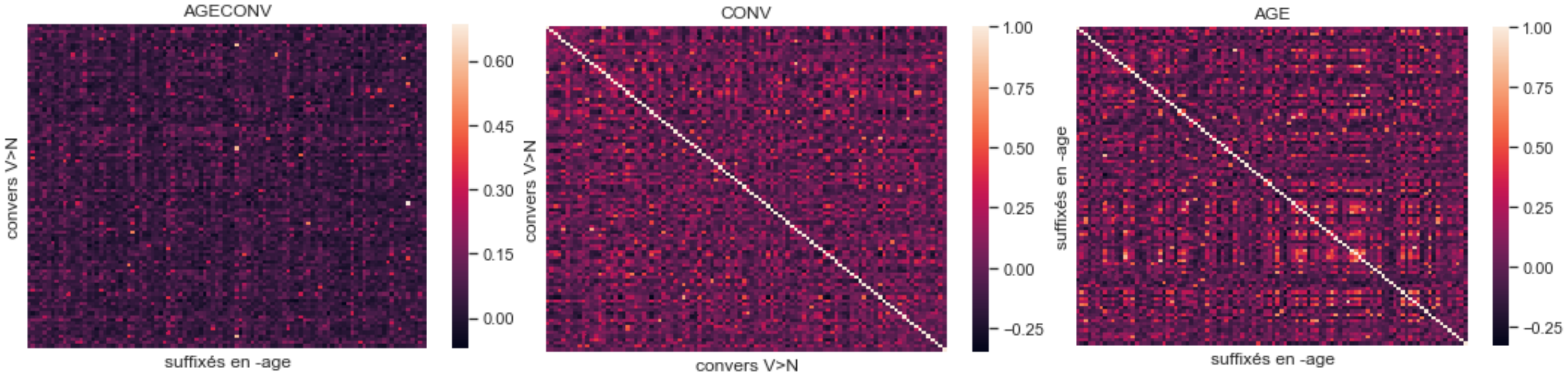
1st experiment: semantic similarities



Dimensionality reduction (t-SNE)

- *-age* derived nouns seem closer to each other than converted nouns.
- Converted nouns seem more scattered.

1st experiment: semantic similarities



Correlation matrices based on cosine similarity:

Less similar → *-age* x conversion
Most similar → *-age* x *-age*

1st experiment: semantic similarities

Are the 10 closest words of a lexeme derived with the same schema?

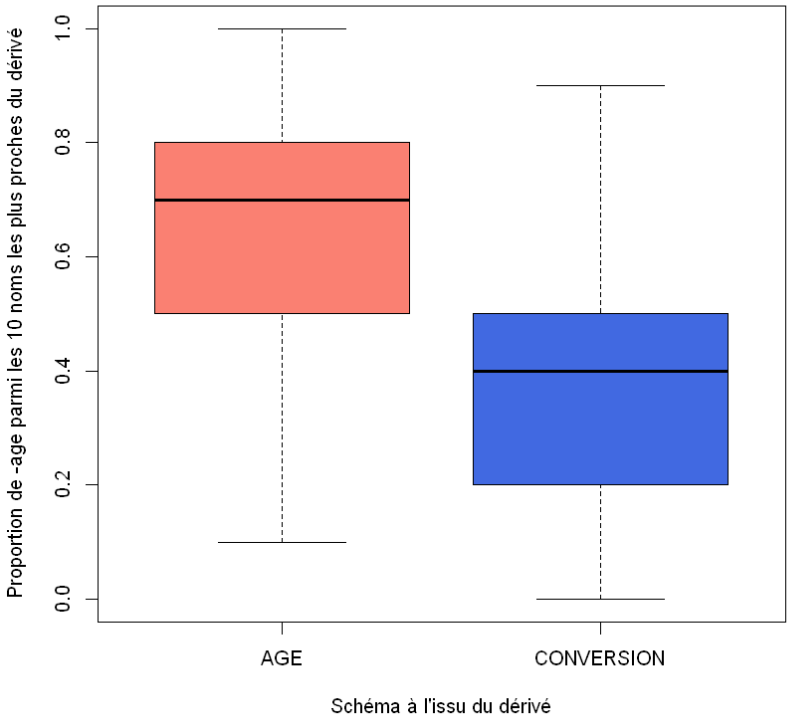
soufflage :

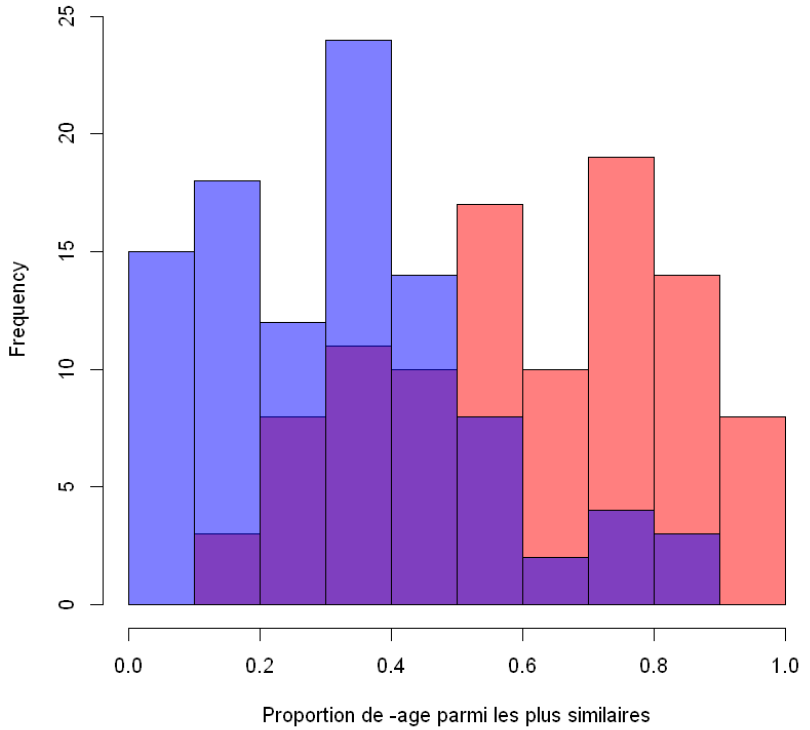
[(’forgeage’, 0.4161148965358734),
(’bouchage’, 0.3646799921989441),
(’puisage’, 0.3522503674030304),
(’vitrage’, 0.3474824428558349 6),
(’émaillage’, 0.3462379276752472),
(’chauffage’, 0.3426979184150696),
(’séchage’, 0.3332732018470764),
(’perçage’, 0.3219649493694305 4),
(’graissage’, 0.2903026044368744),
(’soutirage’, 0.2820050716400146 5)]

1st experiment: semantic similarities

Are the 10 closest words of an –*age* suffixed noun derived with the same schema?

→ On average, **65%** are.

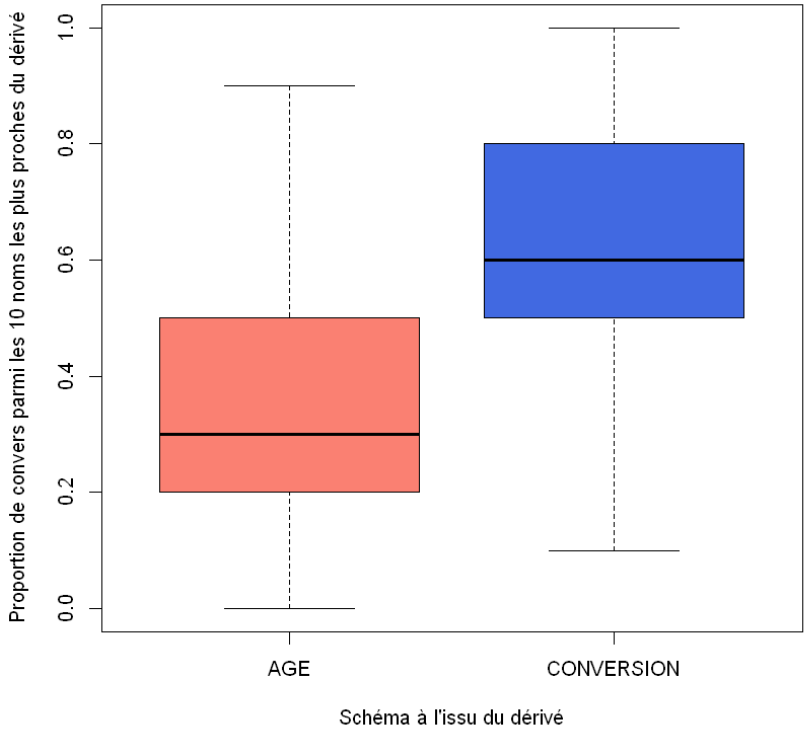


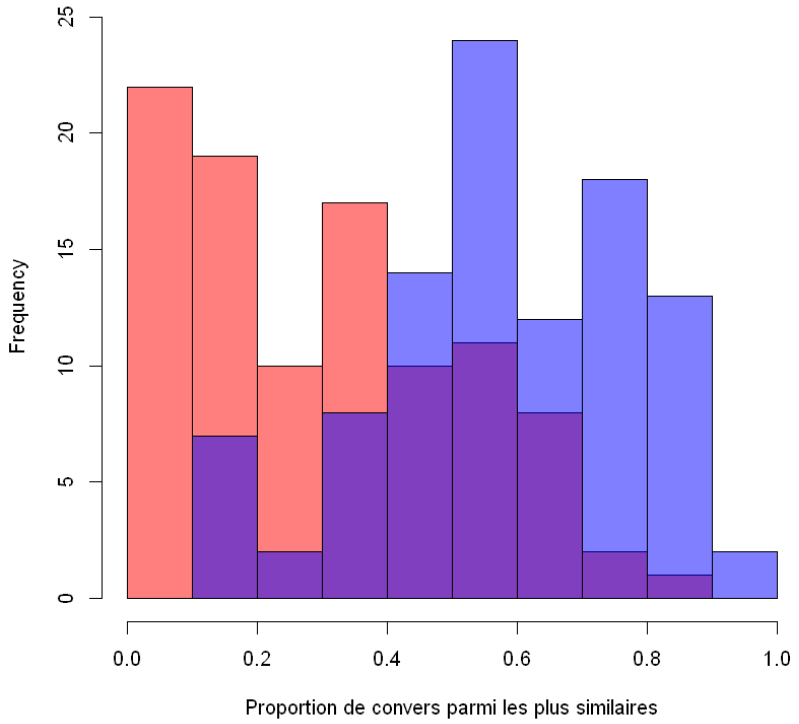


1st experiment: semantic similarities

Are the 10 closest words of a converted noun derived with the same schema?

→ On average, **62%** are.





1st experiment: analysis

What kinds of semantic properties make them distinguishable?

- Converted nouns → none. Semantic types do not matter.
  10 closest nouns can be agent, instrument, event and object nouns. No regularities.

- -*age* suffixed nouns → **activities that need an instrument / an object**.

1st experiment: analysis

What kinds of semantic properties make them distinguishable?

- *-age* suffixed nouns → **activities that need an instrument / an object**.

They represent **65%** of suffixed nouns.

Instruments can sometimes be **morphologically related:**

balayer (avec un balai) → *balayage*
arroser (avec un *arrosoir*) → *arrosage*
chronométrer (avec un chronomètre) → *chronométrage*
percer (avec une perceuse) → *perçage*

Or not:

pirater (avec un ordinateur) → *piratage*
capter (avec un instrument fait pour) → *captage*

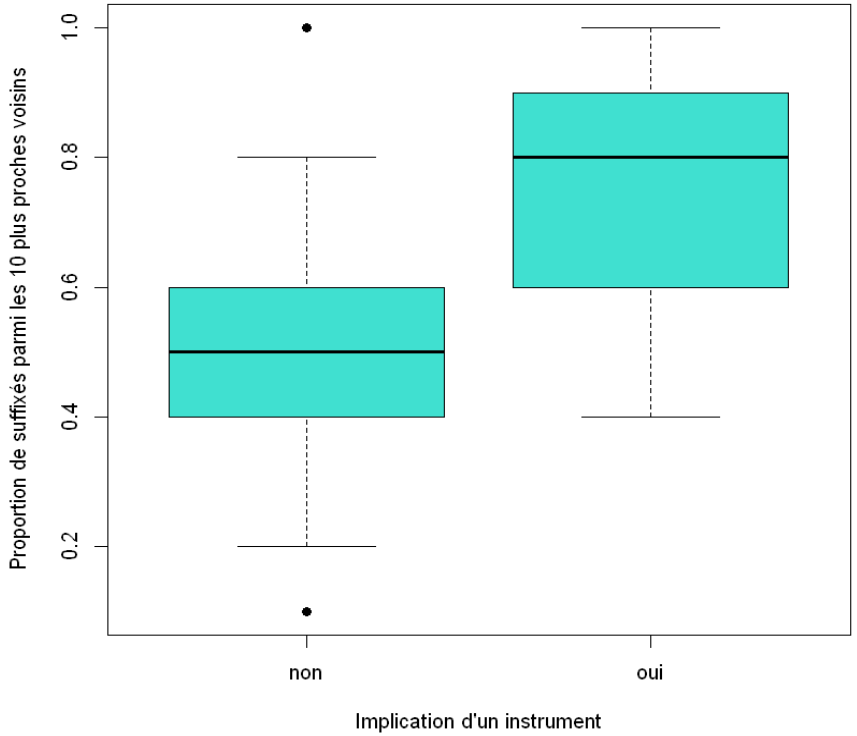
1st experiment: analysis

What kinds of semantic properties make them distinguishable?

- *-age* suffixed nouns → **activities that need an instrument / an object**.

However, only 2 converted nouns denote activities that need an instrument (not morphologically related):

sauvegarder (avec un ordinateur) → *sauvegarde*
retoucher (avec un instrument adéquat) → *retouche*



1st experiment: overview

We found that:

- *-age* suffixed nouns are semantically closer to each other than converted nouns

- *-age* suffixed nouns specialize in activities that need an instrument / an object

→ **A first semantic constraint?**

What about the actual distance between nouns?

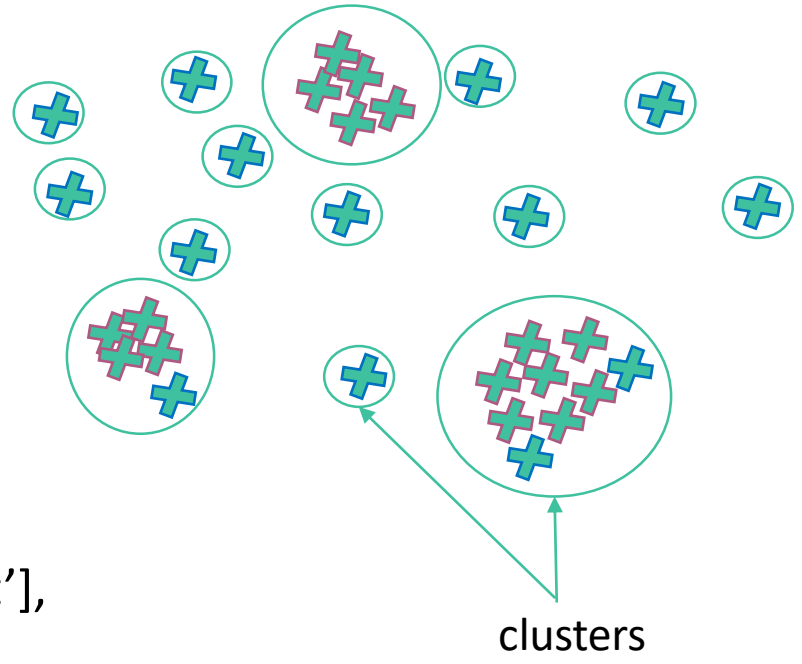
2nd experiment: clustering

K-means algorithm (NLTK):

Generates **clusters of nouns** based on their cosine similarities.

Each noun was assigned to 40 different clusters.
Example with *bronzage* 'suntan':

['éclairage', 'bronzage'],
['mariage', 'maquillage', 'rasage', 'bronzage', 'défilé', 'réconfort'],
['maquillage', 'rasage', 'bronzage', 'défilé'],
['maquillage', 'bronzage'],
['maquillage', 'rasage', 'bronzage']



2nd experiment: proportions among clusters

Quantifying the proportion of –*age* derivatives and converted nouns among clusters for each word.

Example for *bronzage* 'suntan':

['éclairage', 'bronzage'] → age: 1.0, conversion: 0.0
['mariage', 'maquillage', 'rasage', 'bronzage', 'défilé', 'réconfort'] → age: 0.66, conversion: 0.33
['maquillage', 'rasage', 'bronzage', 'défilé'] → age: 0.75, conversion: 0.25
['maquillage', 'bronzage'] → age: 1.0, conversion: 0.0
['maquillage', 'rasage', 'bronzage'] → age: 1.0, conversion: 0.0

Classifier based on the proportions: **predicts a schema for a noun**.

2nd experiment: classification results

Conditions:

- *-AGE* if more clusters of length **>** 2 with a higher proportion of *–age* derivatives
- CONVERSION if more clusters of length **>** 2 with a higher proportion of converted nouns

		ACTUAL	
		-AGE	**CONVERSION**
PREDICTED	**-AGE**	100	98
	CONVERSION	0	2

Predicts 100% of *–age* derivatives and 2% of converted nouns.

When a cluster contains more than 3 nouns, there are more *–age* derivatives

→ *-age* derivatives gather more than converted nouns.

2nd experiment: classification results

Conditions:

- -AGE if more clusters of length **<** 3 with a higher proportion of *–age* derivatives
- CONVERSION if more clusters of length **<** 3 with a higher proportion of converted nouns

		ACTUAL	
		-AGE	**CONVERSION**
PREDICTED	**-AGE**	0	0
	CONVERSION	100	100

Predicts 0% of *–age* derivatives and 100% of converted nouns.

When a cluster contains less than 3 nouns, the exact opposite happens

→ Converted nouns gather in small clusters, stand alone, or join bigger *–age* clusters.

2nd experiment: classification results

Same experiment with nouns without semantic shift from our lexicon:
26 *–age* derivatives, 26 converted nouns

		ACTUAL	
		-AGE	**CONVERSION**
PREDICTED	**-AGE**	26	26
	CONVERSION	0	0

Predicts 100% of –age derivatives and **0%** of converted nouns.

→ Polysemy was not a bias.

2nd experiment: analysis

What do these clusters consist of?

Numerous clusters that denote concrete and specific domains:

Gardening, household tasks, manual labour, beauty care, …

→ Usually activities that need an instrument / object.

AGE clusters

(a) jardinage, arrosage, engrais, désherbage
(b) balayage, repassage, nettoyage
(c) bricolage, brossage, ponçage
(d) marquage, émaillage, séchage, tissage
(e) vitrage, chauffage, ramonage
(f) perçage, brochage, forgeage, pliage, blindage
(g) paramétrage, filtrage, réglage
(h) maquillage, rasage, bronzage, manucure
(i) stockage, entreposage, catalogage, archivage
(j) essayage, arrivage, habillage, tenue, défilé

2nd experiment: analysis

What do these clusters consist of?

Smaller clusters that denote abstract domains:

→ Cooccurents in a specific context (game, aerospatial, economics…)
→ Emotions, damage…

CONVERSION clusters

(a) fusée, poussée, survol
(b) lancée, tournoi, défaite
(c) conglomérat, agrégat, amas
(d) délit, plainte, cambriolage, attentat
(e) revenu, baisse, relance, afflux
(f) mépris, crainte, dégoût
(h) éclaircie, accalmie, secousse

2nd experiment: overview

We found that:

- *-age* suffixed nouns massively invest specific and concrete domains that are better recognized
- Converted nouns are more scattered and invest abstract domains in comparison

→ **Another semantic constraint?**

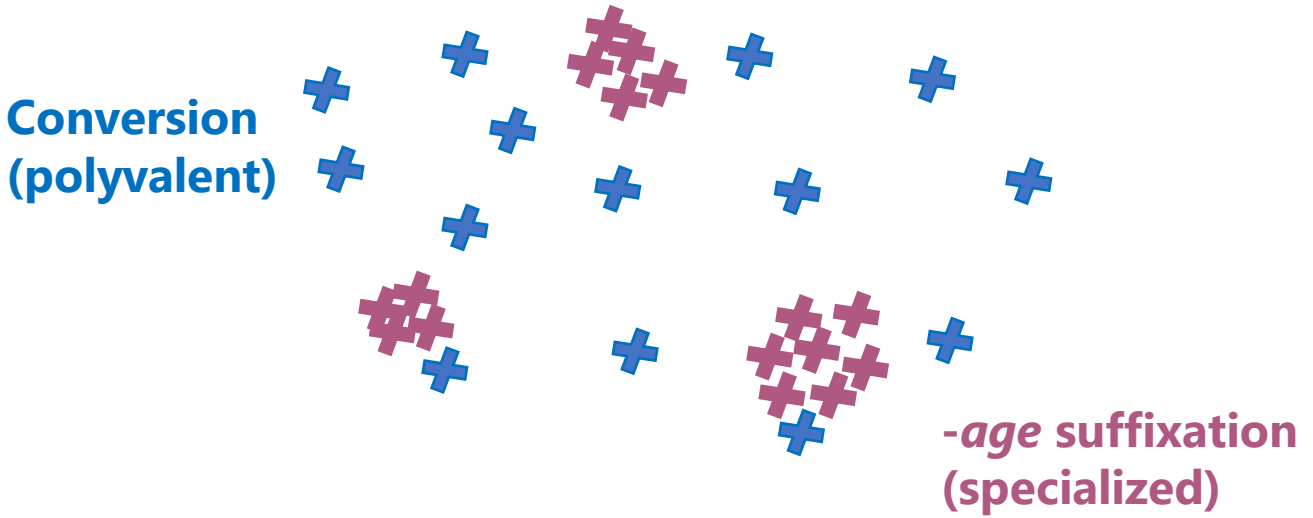
CONCLUSION

- Conversion is polyvalent,
  -*age* suffixation is specialized.

→ Aronoff's morphological niches

- Distributional analysis:
  - Without a priori
  - Coarse-grained analysis of semantic properties

- New perspective on the analysis of morphological rivalry:
  - Interactional constraints within a morphological system
  - Semantic properties organize / drive analogical word-formation



Thank you for your attention!