LexEco: exploring how derivational morphology contributes to the semantics of French nouns

Lucie Barque

Université Sorbonne Paris Nord & LLF

DeriMo

September 4-5, 2025, University of Fribourg, Switzerland

1. Introduction

2. *LexEco*

2.1 Database development

2.2 Statistics

3. Case study

3.1 Simplex vs suffixed nouns: semantic tendencies

3.2 Simplex vs suffixed nouns: ambiguity profiles

4. Conclusion

Introduction

Introduction

- Most word meanings are created by speakers either through morphological processes (1-a) or through polysemous extensions (1-b)

(1) a. *to unfriend* 'remove someone from a list of contacts'
b. *troll* 'a person who provokes others online' (from 'the ugly mythical creature')

- What is the respective contribution of these two mechanisms to the overall economy of meaning production?
- Addressing this question, among others, requires a morpho-semantic description of a representative sample of the lexicon

Introduction

- Comparison between simplex and complex words

	Artifact	**Person**	**Cognition**	**state**	**Attribute**	**Action**
Simplex N	*table*	*mother*	*idea*	*joy*	*charisma*	*embargo*
Complex N	*trawler*	*violinist*	*thought*	*pleasure*	*politeness*	*exhibition*

- Some theoretical studies suggest that morphology plays a complementary role (Croft, 1991)

- Previous empirical studies on French nouns revealed, however, more nuanced patterns (e.g., Tribout et al., 2014; Huyghe et al., 2017; Salvadori, 2024)

LexEco

LexEco

- LexEco is a lexical resource designed to provide a representative sample of the French nominal lexicon (cf. *Echantinom*, Bonami and Tribout (2021)), focusing on the core vocabulary
- Its development is based primarily on existing resources
- Each entry is annotated with morphological, semantic, and both corpus-based frequency and familiarity information

LexEco: Noun selection

- To ensure the ecological validity of the lexicon, nouns were selected from *Lexique 3* (New et al., 2004, 2007), based on familiarity ratings rather than corpus frequency

N	Freq	Fam	N	Freq	Fam
bétel ‘betel’	1.54	30%	tendinite ‘tendinitis’	0.12	100%
gandin ‘dandy’	0.92	25%	peaufinage ‘refinement’	0.1	100%
trèpe ‘huddle’	0.74	19%	physionomiste ‘face reader’	0.1	100%
vertex ‘vertex’	0.61	17%	luxembourgeois ‘Luxembourger’	0.1	100%
boutéon ‘mess tin’	0.57	3%	fluor ‘fluoride’	0.06	100%
voussure ‘arch’	0.41	19%	déforestation ‘deforestation’	0.02	100%

- Nouns with a minimum familiarity of 50% and attested as nouns in the French Wiktionary were retained, resulting in 18,979 nominal lemmas, each associated with textual frequency data (M=16.5, SD=77.7) and familiarity ratings (M=88.5, SD=13.2)

LexEco: Morphological information

- Information on the morphological structure of nouns comes primarily (78%) from four existing morphological resources
  - 2 351 nouns from *Le lexique des noms simples* (Tribout et al., 2014)
  - 3 274 nouns from *Échantinom* (Bonami and Tribout, 2021)
  - 1 513 nouns from *Sonde* (Huyghe et al., sub)
  - 7 760 nouns from *Démonette-2* (Namer et al., 2023)
- The morphological descriptions of the remaining 22% of nouns were produced semi-automatically and partially revised manually
  - Hyphenated nouns in this subset have been automatically classified as compounds
  - Nouns having an adjectival counterpart according to *Lexique-3* have been automatically classified as convert

LexEco: Morphological information

- Information on the morphological structure of nouns comes primarily (78%) from four existing morphological resources
  - 2 351 nouns from *Le lexique des noms simples* (Tribout et al., 2014)
  - 3 274 nouns from *Échantinom* (Bonami and Tribout, 2021)
  - 1 513 nouns from *Sonde* (Huyghe et al., sub)
  - 7 760 nouns from *Démonette-2* (Namer et al., 2023)

- The morphological descriptions of the remaining 22% of nouns were produced semi-automatically and partially revised manually
  - Hyphenated nouns in this subset have been automatically classified as compounds
  - Nouns having an adjectival counterpart according to *Lexique-3* have been automatically classified as convert

LexEco: Morphological information

- Morphological information associated with nouns in *LexEco* adhered to the guidelines established for the construction of *Échantinom* (Bonami and Tribout, 2021)

noun[1]	**cstr**	**suff**	**suff_norm**	**pref**	**conv**	**conv_pos**	**aff_base**	**aff_pos**
cou	simplex	0	0	0	0	0	0	0
embrassade	suffixed	ade	ade	0	0	0	embrasser	V
irrespect	prefixed	0	0	in	0	0	respect	N
réveil	convert	0	0	0	réveiller	V	0	0
cerf-volant	coumpound	0	0	0	0	0	0	0
resto	non-concat.	0	0	0	0	0	0	0
boudeuse	suffixed	euse	eurM	0	boudeur	A	bouder	V
malchanceux	convert	eux	eux	mal	malchanceux	A	chance	N

[1] *cou* ‘neck’, *embrassade* ‘kissing’, *irrespect* 'disrespect’, *réveil* ‘wake-up/alarm clock’, *cerf-volant* ‘kite’, *resto* ‘restaurant’, *boudeuse* ‘a sulky girl/sulky’, ‘insufficiency’

LexEco: Morphological information

- The reliability of the morphological information is still to be assessed
  - For most primary resources, internal consistency of the encoding—reflected by inter-annotator agreement—is not available
  - Diverging approaches to the treatment of complex morphological phenomena, such as
    - Distinction between prefixation and compounding (e.g., *épiphénomène* ‘epiphenomenon’)
    - Suffixation on non-autonomous base (e.g., ablation ‘ablation’)

LexEco: Semantic information

- The semantic information in LexEco is drawn from *SuperWik-fr* (Angleraud et al., 2025), a version of the French Wiktionary in which the senses of ∼230,000 nouns have been automatically annotated with semantic labels

- Word senses are semantically described at two levels of granularity
  - Supersenses (23 classes, e.g., Person, Artifact, Act)
  - Hypersenses (9 classes, e.g., Animate_entity, Inanimate_entity, Dynamic_situation)

(2) LAVE-GLACE ('windshield washer')
a. (Automobile) Dispositif qui envoie du liquide nettoyant sur le pare-brise. '(Automotive) Device that sprays cleaning fluid onto the windshield.' Artifact - Inanimate_entity
b. (Par métonymie) Liquide lave-glace. ex. *Notre antigivre permet de réduire le gel du lave-glace sur le pare-brise, en hiver.* '(By metonymy) Windshield washer fluid. e.g., *Our antifreeze reduces the freezing of the windshield washer on the windshield during winter.*'
Substance - Inanimate_entity

LexEco: Semantic information

- The semantic annotation was performed using supervised classifiers trained and evaluated on a large set of manually curated data
  - Achieved a mean precision of nearly 85% at the supersense level and nearly 92% at the hypersense level
  - Performances vary across semantic categories (F-scores)

Person	**Artifact**	**Act**	...	**Attribute**	**Cognition**	**State**
96.2	86.3	85.9	...	70.4	65.8	62.2

LexEco: Statistics

- Distribution of nouns by types of morphological processes in *LexEco*

	Nb of lemmas	**%**
Suffixation	8,801	46,3
Simplex	5,147	27,1
Conversion	3,645	19,3
Coumpounding	784	4,1
Prefixation	303	1,6
Nonconcat.	304	1,6
	18,984	100

LexEco: Statistics

- Distribution of nominal senses by hypersenses in the dataset[2]

	Nb of senses	**%**
Inanimate_entity	18,945	34
Animate_entity	10,825	19
Dynamic_situation	10,816	19
Stative_situation	5,470	10
Informational_object	5,422	10
Other	5,112	8
	56,590	100

[2] Hypersenses with a representation of less than 3% are grouped under the label `other`.

Case study

Case study: dataset

- Statistics of the dataset reduced to clear-cut[3] cases of simplex and suffixed nouns

	Total N	**Mono. N**	**Ambig. N**	**Senses**	**Mean Ambiguity**	**Freq**
Simplex N	3,971	1,202	2,769	12,802	3.2	28.3
Suffixed N	8,007	2,887	5,120	21,380	2.6	7.1
Total	11,978	4,089	7,889	34,182	2.8	14.1

1. Semantic tendencies among *monosemous* simplex vs suffixed nouns only, as not all senses of ambiguous nouns are morphologically derived[4]
2. Ambiguity profiles of simplex and suffix nouns

[3] Possible cases of conversion were discarded. The number of excluded nouns is higher in the simplex group (956/4,927, 19%) than in the suffixed group (932/8,939, 10%).

[4] (Rainer, 2014; Bauer, 2017; Salvadori, 2024)

Case study: monosemous nouns

Supersense	Hypersense	Simplex		Suffix	
Animal	Animate	8.7	21.3	1.2	29.3
Person		12.6		28.1	
Artifact	Inanimate	16.5	48.6	6.4	12.2
Body		5.1		0.7	
Food		13.1		1.1	
Object		4.7		1.1	
Plant		4.8		1.0	
Substance		4.5		1.9	
Cognition	Information	4.7	5.6	4.3	4.5
Communic.		0.8		0.2	
Act	Dynamic_sit.	7.2	10.1	26.6	32.1
Event		1.7		4.5	
Phenom.		1.2		1.0	
Attribute	Stative_sit.	1.4	4.4	10.1	18.1
Feeling		0.8		1.7	
State		2.2		6.3	
Other (6)	Other (6)	10.1		3.8	

- Simplex nouns mainly denote concrete entities (70%) while suffixed nouns mainly denote abstract entities (58%)
- Within the set of concrete nouns, the balance between animate and inanimate entities is reversed across the two groups
- Within the set of nouns denoting inanimate entities, the balance between artifact and natural objects is reversed across the two groups
- The two groups exhibit significantly distinct semantic distributions ($\chi^2(5, N = 4{,}089) = 845.9$, $p < .001$, Cramer's V $= 0.45$)

Case study: noun ambiguity

- Simplex nouns are significantly more ambiguous[5] (M=3.2) than suffixed nouns (M=2.6), as revealed by a Mann–Whitney U test ($Z = 9.7$, $p < .001$)
- Main, non-exclusive hypotheses
  1. Lexicographic practices, which tend to minimize the number of entries for suffixed N
  2. Frequency: simplex nouns are significantly more frequent than suffixed nouns

     However, the causal relationship between these two collinear variables remains unclear[6]
  3. Semantic specificities of simplex nouns, which mainly denote concrete entities
  4. Lexical longevity, if simplex nouns tend to be older in the lexicon than suffixed forms

[5] Ambiguity is measured by the number of senses attributed to a noun in the French *Wiktionnaire*

[6] (Zipf, 1945; Piantadosi et al., 2012; Koshevoy et al., 2023)

Case study: noun ambiguity

- Poisson regression
  - Dependant variable : number of meanings of N
  - Predictors : log-transformed frequency of N, concreteness of its source meaning

	Estimate	Std Error	z value	Pr($< \|z\|$)
(Intercept)	0.584326	0.010071	58.023	$<$ 2e-16
Concreteness-concrete	0.029294	0.010830	2.705	0.00683
Log_Freq	0.642598	0.007921	81.125	$<$ 2e-16

Case study: ambiguity profile

- The two groups are also expected to show different ambiguity profiles due to:
  1. Their respective semantic tendencies, as observed among monosemous nouns

     Eg. metaphors such as Body→Artifact (e.g., *bouche* ‘mouth/entry’) and metonymies like Body→Person (e.g., *tête* ‘head/intelligent person’) are more typical of simplex Ns
  2. Their respective possible sources of ambiguity
     - For simplex N, ambiguity only results from sense extension
     - For suffixed N, ambiguity results from both sense extension and morphological derivation

- Two broad subtypes of ambiguous words, used as a proxy for their semantic diversity

	Simplex ambigous Ns	Complex ambigous Ns
Monocategorical	TSUNAMI a. *tsunami* Event b. *massive influx* Event	SUFFRAGETTE a. *suffragette* Person b. *feminist* Person
Polycategorical	KEBAB a. hand-held dish Food b. restaurant Institution	CUISINIÈRE a. *female cook* Person b. *kitchen stove* Artifact

Case study: ambiguity profile

- Statistics for the subset of ambiguous nouns

	Lemmas	**Senses**	**Ambiguity**	**Freq**
Simplex-monocat	957	2,717	2.8	29.4
Simplex-polycat	1,812	8,883	4.9	42.9
Suffixed-monocat	2,236	6,107	2.7	7.0
Suffixed-polycat	2,884	12,386	4.2	12.1
Total	7,889	30,093	3.81	19.8

- Monocategorical nouns
  - are significantly more frequent among suffixed than simplex N (43% vs 34%)
  - show no further differences in lexical ambiguity between simplex and suffixed forms, despite displaying comparable differences in frequency
    - Possible effect of morphological derivation

Conclusion

Conclusion

- We presented *LexEco*, a new morpho-semantic lexicon whose key contribution is to provide a representative sample of French nouns known by most adult speakers

- The comparison between suffixed and simplex nouns revealed:
  - A partially complementary distribution of semantic types between the two groups
  - Clear distinctions in ambiguity profiles: simplex nouns are more ambiguous and appear more semantically diverse than suffixed nouns

- Further research:
  - Enhancing the coherence of morphological information in future database releases
  - Conducting more fine-grained semantic analyses of the complementary roles of morphological derivation and polysemy in the construction of nominal meaning

Bibliographie I

Angleraud, N., Barque, L., and Candito, M. (2025). Annotating the french *Wiktionary* with supersenses for large scale lexical analysis: a use case to assess form-meaning relationships within the nominal lexicon. *Proceedings of the 31th International Conference on Computational Linguistics (COLING'2025)*, pages 5321–5332.

Bauer, L. (2017). Metonymy and the semantics of word-formation. In *Mediterranean Morphology Meetings*, volume 11, pages 1–13.

Bonami, O. and Tribout, D. (2021). échantinom: a hand-annotated morphological lexicon of french nouns. In *International Workshop on Resources and Tools for Derivational Morphology*, pages 42–51.

Croft, W. (1991). *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information.* University Press of Chicago.

Huyghe, R., Barque, L., Haas, P., and Tribout, D. (2017). The semantics of underived event nouns in french. *Italian Journal of Linguistics*, 29(1):117–142.

Huyghe, R., Salvadori, J., Varvara, R., Barque, L., Haas, P., Lombard, A., Monney, M., Tribout, D., and Wauquier, M. (sub). Sonde: A database for exploring the semantics of nouns derived from verbs in french. *Morphology*.

Bibliographie II

Koshevoy, A., Dautriche, I., and Morin, O. (2023). Why do some words have more meanings than others? a true neutral model for the meaning-frequency correlation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, pages 2296–2303.

Namer, F., Hathout, N., Amiot, D., Barque, L., Bonami, O., Boyé, G., Calderone, B., Cattini, J., Dal, G., Delaporte, A., Duboisdindien, G., Falaise, A. Grabar, N., Haas, P., Henry, F., Huguin, M., Juniarta, N., Liégeois, L., Lignon, S., Macchi, L., Manucharian, G., Masson, C., Montermini, F., Okinina, N., Sajous, F., Sanacore, D., Tran, T. M., Thuilier, J., Toussaint, Y., and Tribout, D. (2023). Démonette-2, a derivational database for french with broad lexical coverage and fine-grained morphological descriptions. *Lexique*.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661–677.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Rainer, F. (2014). *Polysemy in derivation*, pages 338–353. Oxford University Press.

Bibliographie III

Salvadori, J. (2024). *L'ambiguïté des noms déverbaux en français. Une étude quantitative du sens construit*. PhD thesis, Université de Fribourg.

Tribout, D., Barque, L., Haas, P., and Huyghe, R. (2014). De la simplicité en morphologie. In *SHS web of conferences*, volume 8, pages 1879–1890. EDP Sciences.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2):251–256.