Semantic classes across the lexicon

Nicolas Angleraud, Lucie Barque & Marie Candito

Selexini meeting, 19 June 2024, Université Paris Cité

Plan

1. Introduction

2. Building the semantic resource

3. The resulting resource

4. Semantic classes across the (simplex vs complex) lexicon

1. Introduction

- French lacks comprehensive lexicographic resource that offers semantic information suitable for NLP, like WordNet for English
- Semantic enhancement of the *Wiktionnaire*
  - Semantic classification of nominal senses with supersenses
- Complementary to the FrSemcor, a French corpus annotated in supersenses ($\sim$15,000 occurrences of $\sim$3,000 nouns)

Plan

1. Introduction

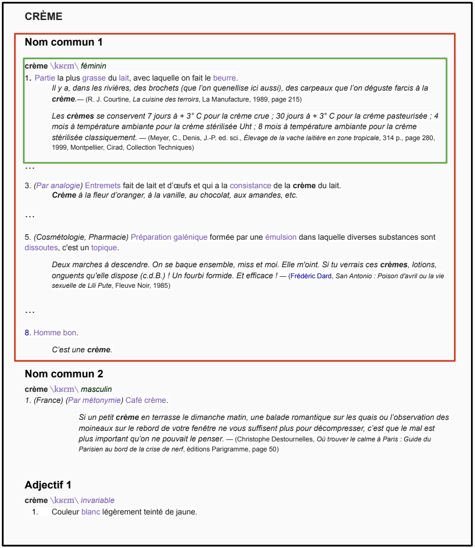
2. Building the semantic resource

3. The resulting resource

4. Semantic classes across the (simplex vs complex) lexicon

2. Building the semantic resource

A *Wiktionary* page



CRÈME

Nom commun 1

crème \kʁɛm\ *féminin*

1. Partie la plus grasse du lait, avec laquelle on fait le beurre.

   Il y a, dans les rivières, des brochets (que l'on quenellise ici aussi), des carpeaux que l'on déguste farcis à la ***crème****.*— (R. J. Courtine, *La cuisine des terroirs*, La Manufacture, 1989, page 215)

   Les ***crèmes*** *se conservent 7 jours à + 3° C pour la crème crue ; 30 jours à + 3° C pour la crème pasteurisée ; 4 mois à température ambiante pour la crème stérilisée Uht ; 8 mois à température ambiante pour la crème stérilisée classiquement.* — (Meyer, C., Denis, J.-P. ed. sci., *Élevage de la vache laitière en zone tropicale*, 314 p., page 280, 1999, Montpellier, Cirad, Collection Techniques)

…

3. *(Par analogie)* Entremets fait de lait et d'œufs et qui a la consistance de la **crème** du lait.

   Crème *à la fleur d'oranger, à la vanille, au chocolat, aux amandes, etc.*

…

5. *(Cosmétologie, Pharmacie)* Préparation galénique formée par une émulsion dans laquelle diverses substances sont dissoutes, c'est un topique.

   Deux marches à descendre. On se baque ensemble, miss et moi. Elle m'oint. Si tu verrais ces ***crèmes****, lotions, onguents qu'elle dispose (c.d.B.) ! Un fourbi formide. Et efficace !* — (Frédéric Dard, *San Antonio : Poison d'avril ou la vie sexuelle de Lili Pute*, Fleuve Noir, 1985)

…

8. Homme bon.

   C'est une ***crème****.*

Nom commun 2

crème \kʁɛm\ *masculin*

1. (France) *(Par métonymie)* Café crème.

> *Si un petit* ***crème*** *en terrasse le dimanche matin, une balade romantique sur les quais ou l'observation des moineaux sur le rebord de votre fenêtre ne vous suffisent plus pour décompresser, c'est que le mal est plus important qu'on ne pouvait le penser.* — (Christophe Destournelles, *Où trouver le calme à Paris : Guide du Parisien au bord de la crise de nerf*, éditions Parigramme, page 50)

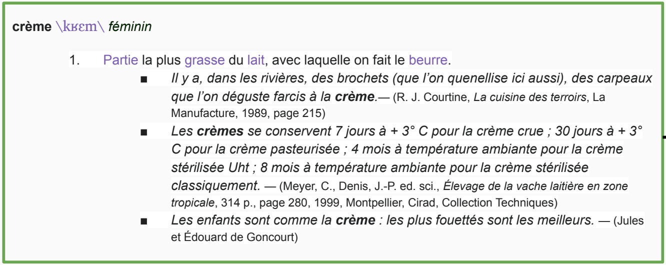
Adjectif 1

crème \kʁɛm\ *invariable*

1. Couleur blanc légèrement teinté de jaune.

2. Building the semantic resource

Objective : Assigning a semantic class to each nominal sense



crème \kʁɛm\ *féminin*

1. Partie la plus grasse du lait, avec laquelle on fait le beurre.
   - *Il y a, dans les rivières, des brochets (que l'on quenellise ici aussi), des carpeaux que l'on déguste farcis à la **crème**.*— (R. J. Courtine, *La cuisine des terroirs*, La Manufacture, 1989, page 215)
   - *Les **crèmes** se conservent 7 jours à + 3° C pour la crème crue ; 30 jours à + 3° C pour la crème pasteurisée ; 4 mois à température ambiante pour la crème stérilisée Uht ; 8 mois à température ambiante pour la crème stérilisée classiquement.* — (Meyer, C., Denis, J.-P. ed. sci., *Élevage de la vache laitière en zone tropicale*, 314 p., page 280, 1999, Montpellier, Cirad, Collection Techniques)
   - *Les enfants sont comme la **crème** : les plus fouettés sont les meilleurs.* — (Jules et Édouard de Goncourt)

→ **Food**

Using :

- The lemma
- The definition
- The examples (when available)

2. Building the semantic resource

Tagset for the semantic classification

Supersenses	**Hypersenses**
Animal, Person	Animate entity
Artifact, Food, Body, Object, Plant, Substance	Inanimate entity
Act, Event, Phenomenon	Dynamic situation
Attribute, State, Feeling, Relation	Stative situation
Cognition, Communication	Informational object
Quantity	Quantification
Institution	Institution
Possession	Possession
Time	Time
Artifact●Cognition	Inanimate entity●Informational object
Act●Cognition	Dynamic situation●Informational object
Group**x**Person	Quantification**x**Animate entity

2. Building the semantic resource

Supervised classification

- Manually annotated data with two subsets of interest

Set	Meanings	Lemmas
train	10,117	4,012
freq-dev	1,581	465
freq-test	1,339	448
rand-dev	540	472
rand-test	649	473
Total	14,226	5,870

- Classifiers
  - Classifier of definition (+lemme)
  - Classifier of occurrences in lexicographic examples
  - Combination of the scores output by the two classifiers
- Models
  - BERT-style pretained language model + Multi Layer Perceptron

2. Building the semantic resource

Results

	rand-dev		freq-dev	
	Supersens	Hypersens	Supersens	Hypersens
frozen bert baseline	61.3	72.8	47.5	57.6
def	78.1	86.5	73.1	78.9
def+lemme	83.3	90.6	76.7	82.2
ex	65.7	77.4	65.0	72.5
def+lemme & ex	**84.3**	**91.3**	**77.1**	**83.0**
monolexical lemmas	85.5	92.3	77.1	83.0
mwe	77.4	85.7	-	-
monosemous lemmas	85.6	91.9	82.5	87.5
polysemous lemmas	77.1	87.9	76.1	82.1

2. Building the semantic resource

Results

Hypersens	Supersens	rand dev		freq dev	
Animate entity	Animal	97.7	90.9	96.6	100.0
	Person		98.4		96.2
Inanimate entity	Artifact	93.9	88.9	90.9	86.3
	Body		76.9		84.9
	Food		76.2		85.7
	Object		57.7		68.4
	Plant		75.0		96 .5
	Substance		75.0		81.4
Dynamic situation	Act	89.2	87.1	86.7	85.9
	Event		75.7		70.0
	Phenomenon		0.0		48.3
Stative situation	Attribute	78.1	83.9	79.7	70.4
	Feeling		85.7		64.0
	Relation				29.6
	State		53.8		62.2
Informational object	Cognition	77.5	66.7	69.9	65.8
	Communication		69.2		74.4
Quantification	Quantity	85.7	85.7	61.2	61.2
Institution	Institution	57.1	57.1	68.1	68.1
Possession	Possession	71.4	71.4	81.8	81.8
Time	Time	66.7	66.7	72.2	72.2

Plan

1. Introduction

2. Building the semantic resource

3. The resulting resource

4. Semantic classes across the (simplex vs complex) lexicon

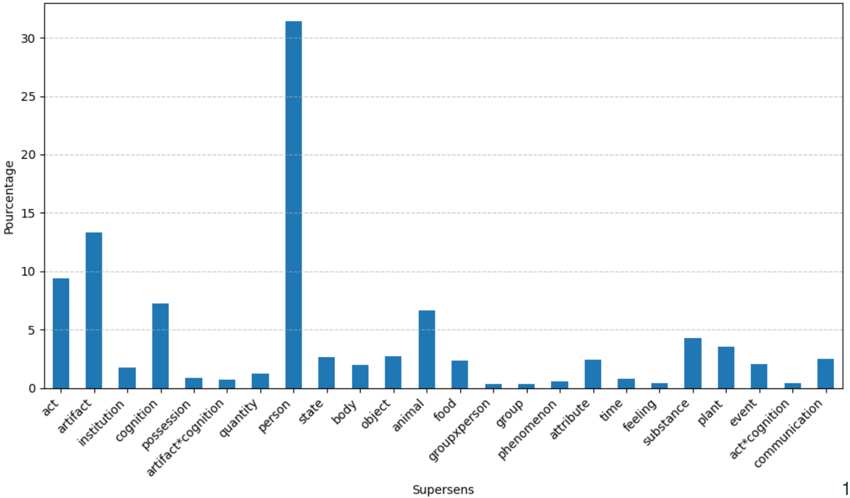
3. The resulting resource

Statistics

Information	Value
Number of meanings	306,225
Number of lemmas	228,989
Ratio nb of meanings / nb of lemmas	1.34
Lemmas with several entries (homonymy)	2%
Proportion of monosemous lemmas	83%
Proportion of MWEs	20%
Proportion of meanings with no lexicographic example	50%
Proportion of demonyms	20%

3. The resulting resource

Distribution of supersenses in the lexical resource



3. The resulting resource

Applications

- Quantitative semantic analysis to answer linguistic questions
    - Eg. form-meaning relationships (cf next section)
- External knowledge to improve semantic nlp tasks
    - Helping supersense tagging in context, especially for rare nouns
    - Verbal disambiguation
- Further semantic enhancement of the resource
    - Meaning frequency estimates, using semantically annotated monosemous words (cf. methodology adopted in (Aloui et al 2021))
    - Partial hierarchy of meanings

Plan

1. Introduction

2. Building the semantic resource

3. The resulting resource

4. Semantic classes across the (simplex vs complex) lexicon

4. Semantic classes across the (simplex vs complex) lexicon

A first assessment of Croft's hypotheses

- Simple morphological nouns prototypically denote objects, while action nouns are prototypically constructed from verbs and property nouns from adjectives (Croft, 1991, 2022)
- A first assessment on 3,489 manually annotated simple nouns (Tribout *et al.* 2014)
  - $\sim$ 75% of the annotated nouns denote Object
  - A tripartite classification not comprehensive enough (eg. *mardi* 'tuesday')
  - Polysemy has to be taken into account (eg. *bœuf* 'jam session')
- WikClasSem allows for a wider empirical investigation of Croft's hypothesis on the French lexicon

4. Semantic classes across the (simplex vs complex) lexicon

Selection of the dataset

- External resources
  - Frequency information
    - *Lexique 3* (New et al. 2004)

      ⇒ Unbiased sample
  - Morphological information
    - *Demonette-2* (Namer *et al.* 2023)
    - *Echantinom* (Bonami & Tribout 2021)

      ⇒ Morphological process, base POS

- New subset :

  (Demonette-2 ∪ Echantinom) ∩ Lexique-3 ∩ WikClasSem

  ⇒ 17,474 nouns

  ⇒ 47,500 nominal meanings

4. Semantic classes across the (simplex vs complex) lexicon

Types of word formation processes

	Nb of lemmas	Proportion	Example
Suffix	9,032	51.7%	*cotisation*
Simplex	5,488	31.4%	*heure*
Conversion	2,476	14.2%	*siège*
Polylexical	271	1.6%	*hors-bord*
Nonconcat	115	0.7%	*micro*
Prefix	80	0.3%	*reflux*
Pre-suf	11	0.1%	*coreligionnaire*
Total	17,474		

4. Semantic classes across the (simplex vs complex) lexicon

Distribution of meanings (in hypersenses)

	Nb of meanings	Proportion
animate	3888	23.3%
dynamic_situation	4193	24.0%
dyn_sit●info	207	1.2%
inanimate	4891	28.0%
inanimate●info	114	0.7%
info	1472	8.4%
institution	251	1.4%
possession	122	0.7%
quantification	178	1.0%
quanti*x*animate	67	0.4%
stative_situation	1967	11.3%
time	123	0.7%
Total	17,473	

4. Semantic classes across the (simplex vs complex) lexicon

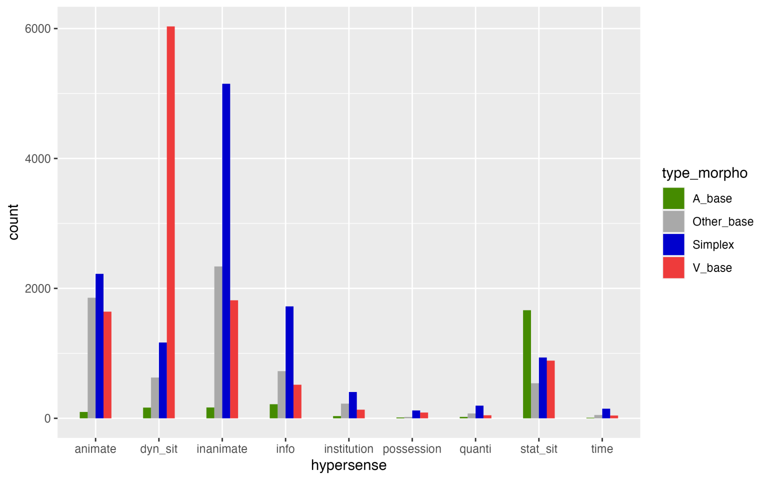
Questions

- What proportion of simplex nouns denote object ?
- To what extent do action nouns derive from verbs ?
- To what extent do property nouns derive from adjectives ?

⇒ Focus on simplex and suffixed nouns

4. Semantic classes across the (simplex vs complex) lexicon

Sense distribution (for simplex, V-base and A-base nouns)



4. Semantic classes across the (simplex vs complex) lexicon

Refining Croft's predictions

- Simplex nouns
  - Mostly denote object (**61%**)
  - But also
    - abstract entities (21.5%)
    - dynamic or stative situations (17.5%)
- Action nouns
  - Are mostly derived from verbs (**75.5%**)
  - But also
    - from other POS (10%)
    - or correspond to simplex nouns (14.5%)
- Property nouns
  - Are frequently derived from adjective (**41%**)
  - But not mostly
    - Simplex (23.2%), V_base (22.0%), other bases (13.4%)

Bibliographical references

Aloui, C., Ramisch, C., Nasr, A., & Barque, L. (2020, December). SLICE : Supersense-based lightweight interpretable contextual embeddings. In The 28th International Conference on Computational Linguistics (COLING 2020). Barque, L., Haas, P., Huyghe, R., Tribout, D., Candito,

M., Crabbé, B. et Segonne, V. (2020). FrSemCor : Annotating a French corpus with supersenses. *LREC-2020*, May 2020, Marseille, France.

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

Croft, W. (2022). Morphosyntax : constructions of the world's languages. Cambridge University Press.

Namer, F., Hathout, N., Amiot D. et al. (2023) Démonette-2, a derivational database for French with broad lexical coverage and fine-grained morphological descriptions, *Lexique*, 23.

New, B., Pallier, C., Brysbaert, M., Ferrand, L. (2004) Lexique 2 : A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers*, 36 (3), 516-524.

Tribout, D., Barque, L., Haas, P., & Huyghe, R. (2014). De la simplicité en morphologie. In SHS web of conferences (Vol. 8, pp. 1879-1890). EDP Sciences.