What do low frequency lexemes tell us ?
A study of the competition in verb formation in French

Fabio Montermini & Juliette Thuilier

Université de Toulouse Jean Jaurès & CLLE-ERSS

JE Concurrence & Polysémie, June 07 2019

Introduction: The empirical domain

Competition between morphological processes in the formation of verbs from nouns or adjectives in French (Roger, 2003; Tribout, 2010; Lignon, 2013; Bonami and Thuilier, 2018)

	Nominal base	Adjectival base
-iser	CARAMEL ~ CARAMÉLISER	CONCRET ~ CONCRÉTISER
	'caramel' ~ 'caramelize'	*'concrete' ~ 'concretize'*
-ifier	MUSÉE ~ MUSÉIFIER	CLAIR ~ CLARIFIER
	'museum' ~ 'transform into a museum'	*'clear' ~ 'clarify'*
en-	PILE ~ EMPILER	RICHE ~ ENRICHIR
	pile ~ 'pile up'	*'rich' ~ 'make rich'*
a-	LIGNE ~ ALIGNER	PLAT ~ APLATIR
	'line' ~ 'line up'	*'flat' ~ 'flatten'*
conversion	COURBE ~ COURBER	MÛR ~ MÛRIR
	curve ~ 'curve'	*'rip' ~ 'ripen'*

- Exclusion of processes implying a negative or reversative reading (*é-*, *dé-*)
- Which (formal or semantic) factors correlate with the choice of the process?

Introduction: Which data?

- Very low frequency lexemes
  - The chances of accessing lexemes created 'on the spot' by speakers are higher among rare words, even if not all low-frequency lexemes are nonce creations.
    - About 20% of the items in our dataset display 'occasionalness' markers (quotes, periphrases, metalinguistic comments; Dal and Namer 2016);
  - less influenced by lexicalisation, semantic shifts, loss of transparency, polysemy...;
  - reasonable size of the dataset for fine-grained manual (especially syntactico-semantic) annotation;
  - syntactico-semantic annotation of real occurrences, rather than of abstract lexemes.

Outline

The dataset: Extraction of the data

- Construted verbs having frequency 1 in FrWac (Baroni et al. 2009)
- Automatic extraction of forms lemmatized as *verb* displaying the target affixes
- Automatic extraction of converted verbs on the basis of a lexicon of N / A (GLÀFF, Sajous et al. 2013)
- Manual cleaning (86% left out)
- 658 items overall.

conversion	*en-*	*a-*	*-ifier*	*-iser*
206 (31.3%)	28 (4.5%)	0 (0%)	48 (7.3%)	376 (57.1%)

  - Only figures for prefixed and suffixed verbs are comparable (not conversions).

The dataset : First annotations

- Phonological properties:
  - length of the stem in number of syllables (the *radical* in the sense of Roché 2010),
  - first and last phonemes of the derivational stem ;
- Morphological properties:
  - The problem : which lexeme should be considered the base of the verb ?

(1) $\text{GAY}_{N/A} > \text{GAYIFIER}_V$

'*become a gay/make gay*'

in "*Zacks qui souffre quand même de la tendance à* ***Gayifier*** *tous ces persos de Nomura.* "

(2) $\text{MÈRE}_N/\text{MATERNEL}_A > \text{MATERNALISER}_V$

'*act like a mother/ to mother*'

in "*De plus, elle a été enchantés, lui permettant de la* ***maternaliser*** *à volonté, au prix d'un depense d'énergie relativement importante.*"

Ascending Morphological Family

The *Ascending Morphological Family* (AMF, Bonami and Thuilier 2018) corresponds to the relevant area of the morphological family containing any lexeme considered as a possible ancestor of the derived verb.

Class	Noun	Adjective	Constructed verb
N	AUTOPORTRAIT	–	AUTOPORTRAITER
A	–	CRADE	CRADIFIER
both	TOURISTE	TOURISTIQUE	TOURISTICISER

Table: The AMF variable

1. Introduction

2. Dataset

3. First results
   - Comparison with Bonami and Thuilier's (2018) data
   - Competition between the processes

4. Syntactico-semantic properties
   - Annotation
   - Results
   - Perspectives

5. Conclusion

Comparison of our data and Bonami and Thuilier's

- Bonami and Thuilier (2018) study the rivalry between *-iser* and *-ifier* suffixes, on the basis of data extracted from a large-scale lexicon (GLÀFF, Sajous et al. 2013)
- Do frequency 1 lexemes behave similarly?

- Our *-iser* / *-ifier* data

Freq. 1 lexemes	
-ifier	*-iser*
48	376
11.3%	*88.7%*

- Bonami and Thuilier's data

Lexicon data	
-ifier	*-iser*
92	699
11.6%	*88.4%*

Comparison of our data and Bonami and Thuilier's

- Comparable effect of the stem length

	Freq. 1 lexemes		Lexicon data	
	-ifier	*-iser*	*-ifier*	*-iser*
1 syll	25 (53.2%)	22 (46.8%)	60 (62.5%)	36 (37.5%)
2 syll	19 (9.5%)	182 (90.5%)	31 (8.1%)	350 (91.8%)
more than 2	4 (2.3%)	172 (97.7%)	1 (0.3%)	313 (99.7%)

Comparison of our data and Bonami and Thuilier's

- Comparable effect of the last consonant of stems
  1. Alveolar obstruents [t, d, s, z] slightly disfavor the *-iser* suffix, complying with a dissimilative constraint
  2. Stems ending in sonorants [l, r, m, n] are more likely to be combined with the *-iser* suffix

	Freq. 1 lexemes		Lexicon data	
	-ifier	*-iser*	*-ifier*	*-iser*
Alveolar Obst.	23 (21.5%)	84 (78.5%)	41 (23.8%)	131 (76.2%)
Sonorant	17 (6.5%)	245 (93.5%)	38 (6.8%)	518 (93.2%)
Other	8 (14.5%)	47 (85.5%)	13 (20.6%)	50 (79.4%)

Comparison of our data and Bonami and Thuilier's

- Effect of the AMF

	Freq. 1 lexemes		Lexicon data	
	-ifier	*-iser*	*-ifier*	*-iser*
Only A	13 (22%)	46 (78%)	13 (21.7%)	47 (78.3%)
Both	22 (10.5%)	188 (89.5%)	59 (8.1%)	558 (90.4%)
Only N	13 (8.4%)	142 (91.6%)	20 (17.5%)	94 (82.5%)

- In both datasets,
  - AMFs containing only a candidate adjectival base sligthly favor *-ifier* suffixation
  - AMFs containing both types of candidate bases favor *-iser* suffixation
- However, AMFs containing only a candidate nominal base favor *-ifier* suffixation according to Bonami and Thuilier's dataset, whereas it favors *-iser* suffixation according to ours.

Comparison of our data and Bonami and Thuilier's

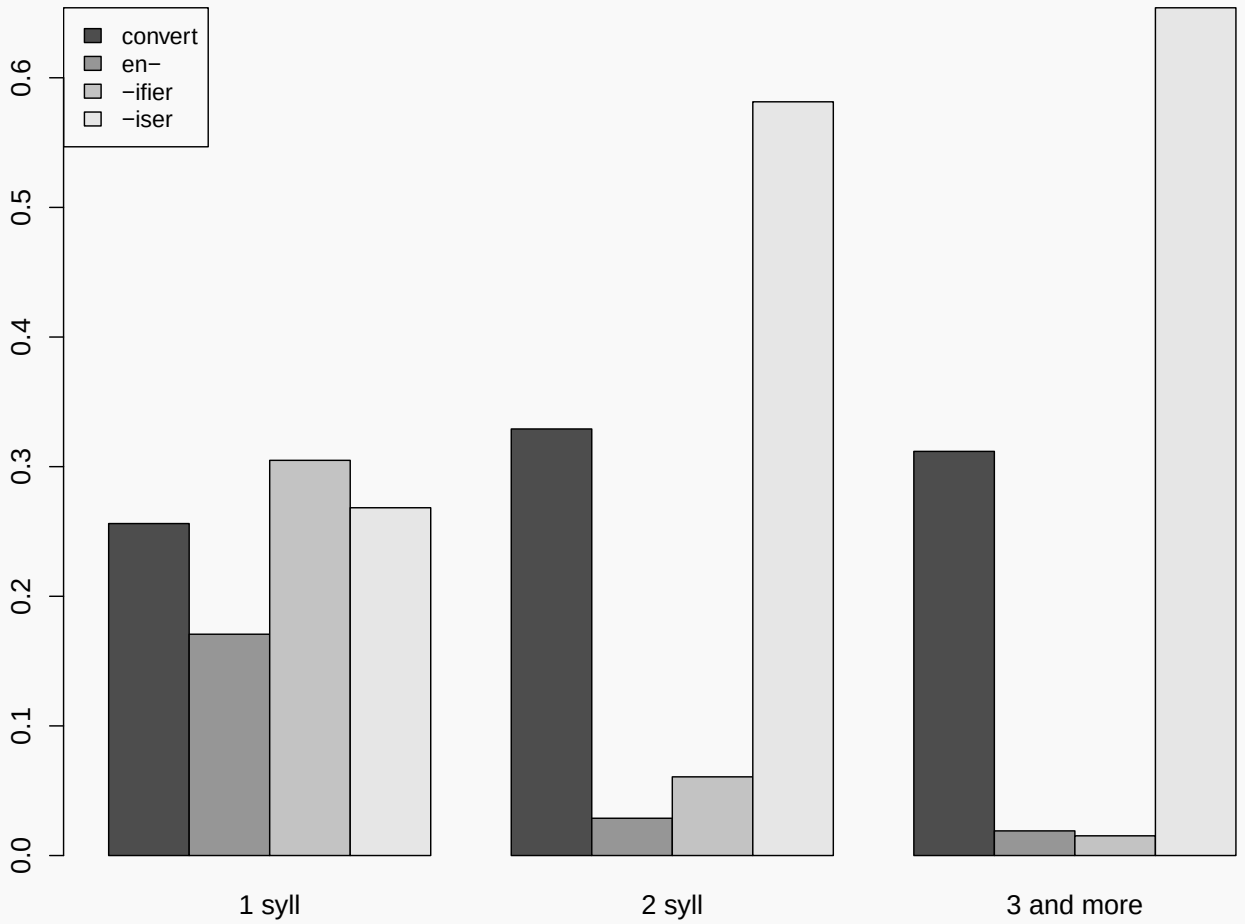
- These observations are confirmed by statistical modelling

⇒ From the formal point of view, there seems to be no difference between frequency 1 data and large-scale databases.

⇒ With regard to AMF, the difference observed may be due either to a difference in annotation or to a real difference in the functioning of *-iser* and *-ifier* in synchrony.

Competition between the 4 processes : Phonological properties

- Morphological competition is conditioned by the length of the stem.



Competition between the 4 processes : Phonological properties

- As mentioned previously, the properties of the last phoneme of the stem conditioned the choice of the suffix.
- Morphological competition is also conditioned by properties of the first phoneme of the stem.
  1. When the stem begins with a vowel, *en-* prefix is dispreferred

	conversion	*en-*	*-ifier*	*-iser*
Consonant	167 (31.2%)	28(5.2%)	39 (7.3%)	302(56.3%)
Vowel	39 (32%)	0	9 (7.4%)	74 (60.7%)
Total	206 (31.3%)	28 (4.5%)	48 (7.3%)	376 (57.1%)

Competition between the 4 processes : Morphological properties

- The shape of the morphological family seems to affect the choice of the morphological strategy
  - AMFs with only A bases favor *-ifier* suffixation
  - AMFs with both bases favor *-iser* suffixation
  - AMFs with only N bases favor conversion

	conversion	*en-*	*-ifier*	*-iser*	Total
only A	11 (15.5%)	1 (1.4%)	13 (18.3%)	46(64.8%)	71 (100%)
both	45 (17.2%)	6 (2.3%)	22 (8.4%)	188 (72%)	261 (100%)
only N	150 (46%)	21 (6.4%)	13 (4%)	42 (43.6%)	326 (100%)
Total	206 (31.3%)	28 (4.5%)	48 (7.3%)	376 (57.1%)	658(100%)

1. Introduction
2. Dataset
3. First results
   - Comparison with Bonami and Thuilier's (2018) data
   - Competition between the processes
4. Syntactico-semantic properties
   - Annotation
   - Results
   - Perspectives
5. Conclusion

Syntactico-semantic annotation

- Inter-annotator agreement on a sample of 329 items:
  - Transitivity: 83,4%;
  - Telicity: 86%;
  - Semantic role of the base: 89,7%;

Syntactic annotation

- Transitivity of the derived verb in its specific context (transitive / pronominal / intransitive);
- Restrictive view (for practical reasons): considered as transitive only in presence of an explicit object NP / pronoun (see below).

Semantic annotation

- Semantic classification (Plag, 1999)

Locative	put into X	HOSPITALIZE	suff. + conv.
Ornative	provide with X	PATINIZE	suff. + conv.
Causative	make (more) X	RANDOMIZE	suff. + conv.
Resultative	make into X	PEASANTIZE	suff. + conv.
Inchoative	become X	AEROSOLIZE	suff. + conv.
Performative	perform X	ANTHROPOLOGIZE	suff. + conv.
Similative	act like X	POWELLIZE	suff. + conv.
Instrumental	use X	HAMMER	conv.
Privative	remove X	BARK	conv.
Stative	be X	HOSTESS	conv.

- This classification (and its outcomes, e.g. Namer 2013; Bonami and Thuilier 2018) cross-cuts different properties of verbs, including aspect and valency.

Semantic annotation

	Telic	Atelic
Predicate	ABSOLUTIFIER Caus./Res./Inc.	FESTIVALER Stative/Perf.
Predicate (indirect)	HOMÉOPATHISER Performative	OFFSHORISER Performative
Locative	FOURRIÉRISER Loc./Orn.	CANAPISER
Agent	BENLADENISER	MATERNALISER Similative
Instrument	BISTOURISER	MARTEAUPIQUEURISER Instrumental

Semantic annotation

- Here, a (provisional) annotation of the role of the base lexeme that takes into account the event denoted by the derived verb and its frame:
  - the base lexeme denotes the event itself (process or endpoint) > Predicate;
  - the base lexeme denotes an element within the frame (typically an agent or an instrument) > Participant;
  - the base lexeme denotes both the event (typically its endpoint) and an element within the frame > Locative.

Event		*Predicate*
Event	Participant	*Locative*
	Participant	*Participant*

- Perspective: finer semantic annotation.

Semantic properties

- Locative meaning favors *en-* prefixation

	conversion	*en-*	*-ifier*	*-iser*
Participant	36 (54.5%)	1 (1.%)	1 (1.5%)	28 (42.4%)
Locative	20 (36.4%)	13 (23.6%)	3 (5.5%)	19 (34.5%)
Predicate	146 (27.6%)	5 (1.4%)	34 (9.6%)	218 (61.4%)
Total	206 (31.3%)	28 (4.5%)	48 (7.3%)	376 (57.1%)

Semantic properties

- Atelic verbs are mainly derived by conversion
- Telic verbs are mainly derived by suffixation

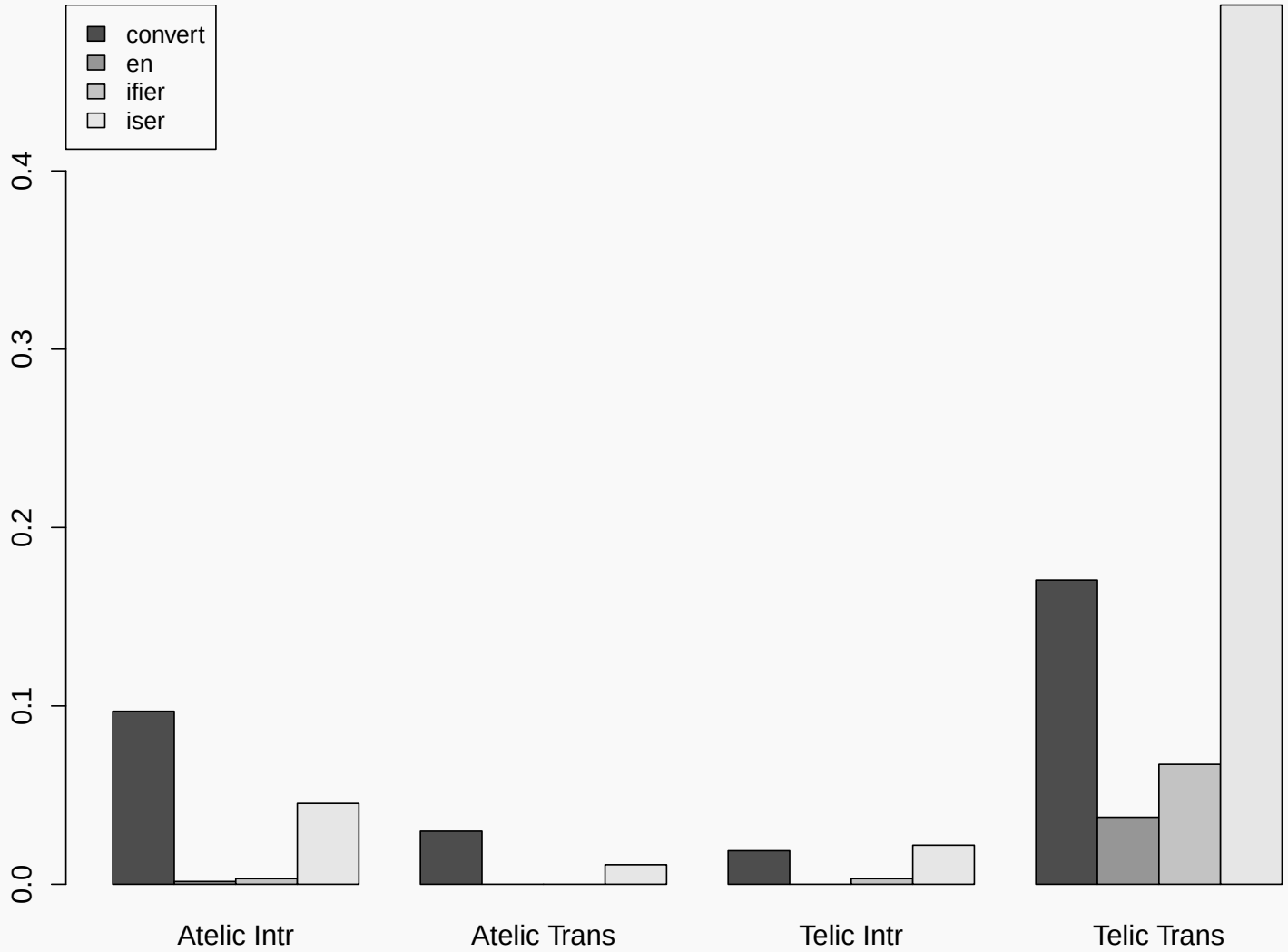
	conversion	*en-*	*-ifier*	*-iser*
Atelic	81 (67.5%)	1 (0.8%)	2 (1.7%)	30 (30%)
Telic	121 (23.3%)	24 (4.6%)	45 (8.7%)	329 (63.4%)
Total	206 (31.3%)	28 (4.5%)	48 (7.3%)	376 (57.1%)

Syntactic properties

- Transitivity favors suffixation and intransitivity favors conversion;
- Same distribution as for telic / atelic verbs:

	conversion	*en-*	*-ifier*	*-iser*
Intransitive	74 (60.3%)	1 (0.8%)	4 (3.3%)	43 (35.5%)
Trans./Pron.	128 (24.8%)	24 (4.6%)	43 (8.3%)	322 (62.3%)
Total	206 (31.3%)	28 (4.5%)	48 (7.3%)	376 (57.1%)

Telicity / transitivity



Marginal cases

- Telic / intransitive:
  - Often, implicit object, recoverable from the context:

    (3) "*Je me demande qui donne l'ordre de* ***pévétiser*** *et* ***fourriériser*** *immédiatement ou de faire l'effort de contacter les propriétaires.*"

Marginal cases

- Atelic / transitive:

  (4) "*Ne l'obligeons donc pas, ivre de désespoir, à* ***"marteaupiqueuriser"*** *et* ***"tractopelliser"****son oeuvre.*"

  (5) "*Je suis bien à* ***nariner*** *les pistils du printemps qui, malgré cette campagne achevée toute en douleur, m'apaisent.*"

- Interestingly, in most cases the base corresponds to a participant (agent or instrument).

Marginal cases

- To be distinguished from 'generic' predicates, which mostly correspond to a plurality of telic events, and may be transitive or intransitive:

(6) "*Dans les années soixante,* ***"stéréophoniser"*** *des enregistrements mono en utilisant des filtres en peigne était un sport à la mode.*"

(7) "*La réalité, la vraie, est parfois un peu trop nuancée pour frapper les esprits. Il faut exagérer,* ***manichéaniser****, amplifier!.*"

Perspectives: problems

- A finer semantic annotation:
- e.g. "Indirect Predicates":

(8) "*La seule solution viable du point de vue du développement durable est d'**électroniser** la conduite automobile en régulant le trafic à l'aide des procédés modernes que sont le GPS et la communication RFiD.*"

(9) "*il aurait pu **télépathiser** à sa femme l'événement en termes assez clairs pour atteindre son but.*"

Perspectives: problems

- How to deal with the overabundance induced by the choice of using frequency 1 lexemes?
  - Several cases of 'affix' or 'stem substitution':

  (10) ACTIVISER, COMMENTARISER, CONFUSER, EXHIBITIONNER, POURRIFIER, SYMBOLIFIER

Perspectives: cross-linguistic comparison

- Italian has a quite similar distribution of constructions creating verbs from nouns / adjectives (conversion, *-izzare*, *-ificare*, *a-*, *in-*);
- Unlike French, it also possesses two constructions which mostly (but not exclusively) derive atelic verbs (*-eggiare*, *s-*);
- Which are the possible consequences on the global equilibrium of the morphological system?

Conclusion

- A global account of denominal / deadjectival verbal derivation:
  - 4 processes involving different morphological strategies;
  - account for the intertwining of formal, syntactic and semantic factors;
  - statistical modelling of the competition.

- As far as formal properties are concerned, our (preliminary) results on frequency 1 lexemes show no significant discordance with analyses performed on larger databases that include frequent items;

Conclusion

- Low frequency (and in particular frequency 1) lexemes allow a fine-grained semantic analysis of derived lexemes (and of the semantic role played by their bases);
- This implies, among other things, a change in perspective: since low frequency derivatives are likely to correspond to nonce forms occasionally created by speakers, semantic analyses are not made on abstract lexemes, but on concrete occurrences, which are, nevertheless, manifestations of the morphological competence speakers have.

Thank you!

References I

Baroni, M., Barnardini, S., Ferraresi, A., and Zanchetta, E. (2009). 'The wacky wide web : a collection of very large linguistically processed web-crawled corpora'. *Language Resources and Evaluation*, 43:209226.

Bonami, O. and Thuilier, J. (2018). 'A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*'. *Word Structure*, 11.

Dal, G. and Namer, F. (2016). 'À propos des occasionnalismes'. In *Actes du Congrès Mondial de Linguistique Française*. Tours, 1–18.

Lignon, S. (2013). '-ISER and -IFIER suffixation in French: Verifying data to 'verize hypotheses''. In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse, 2-3 December 2010)*. Munich: Lincom Europa, 119–132.

References II

Namer, F. (2013). 'Adjectival bases of french -aliser and -ariser verbs: Syncretism or under-specification?' In N. Hathout, F. Montermini, and J. Tseng (eds.), *Morphology in Toulouse. Selected Proceedings of Décembrettes 7 (Toulouse, 2-3 December 2010)*. München: Lincom Europa, 83–102.

Plag, I. (1999). *Morphological productivity*. Berlin: Mouton de Gruyter.

Roché, M. (2010). 'Base, thème, radical'. *Recherches linguistiques de Vincennes*, 39:95–134.

Roger, C. (2003). 'Derived change-of-state verbs in french: a case of semantic equivalence between prefixes and suffixes'. *Acta Linguistica Hungarica*, 1-2:187–199.

References III

Sajous, F., Hathout, N., and Calderone, B. (2013). 'GLÀFF, un Gros Lexique À tout Faire du Français'. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*. Les Sables dOlonne, 285298.

Tribout, D. (2010). *La conversion de nom à verbe et de verbe à nom en français*. Ph.D. thesis, Université Paris Diderot (Paris 7).