Néoveille - An automatic System for Lexical Units Life-Cycle Tracking

Emmanuel Cartier

Université Paris 13 Sorbonne Paris Cité, LIPN-RCLN CNRS UMR 7030, Labex EFL

Euralex 2018, Ljubljana, July 19th 2018

MOTIVATION

- **Computational linguistics :**
  - **modelize lexical change so as to be able to detect it automatically on a large scale**
  - **provide an online tool for linguists to (un)validate these automatically detected changes, describe them and track their life-cycle**
- **Linguistics and lexicography :**
  - **better understanding of the functioning of lexical change : formal and semantic mechanisms at stake, lifecycle models, etc.**
  - **obtain a sketch of the lexicon trends for a given language**

CONTENTS

A. (Quick) Theoretical assumptions

B. Néoveille architecture and modules

C. Quantitative and Qualitative Results for French

D. Conclusions and perspectives

(TOO QUICK) THEORETICAL ASSUMPTIONS

Evidence of (continuous) lexical change : in discourse, about 5 % of lexical units are outside the scope of dictionary coverage (Renouf, 2014; Cartier, 2016)

Intuitive definition of neologism (or lexical innovation, LI) : any lexical item or usage deviating from the assumed usage of the speech community. From the first occurrence in corpora.

Dynamics of language (Coseriu, 1954), (Weinreich and al., 1968) :

- revisiting Saussure dichotomy Langue / discours : discourse enables the preservation of the language system, but at the same time continuously modifies it by introducing new lexical items or new usage of existing lexical items, and application to new referents
- adding a pre-variationist point of view : a lexical change occurs in a specific speech community - and thus is first a variation - and (can) diffuse through several speech community before being adopted by the whole community.

Usage-based linguistics : collocations (Firth, 1957), *collostructions* (Stefanowitsch et Gries, 2003), *collocational profile* (Sinclair, 1991), *profil combinatoire* (Blumenthal, 2005) or *behavioral profile* (Gries, 2010)

Cognitive linguistics / Construction Grammars : from linguistic sign to construction (Goldberg, 2013), constructionalization (Traugott and Trousdale, 2013) and entrenchment (Langacker, 1990)

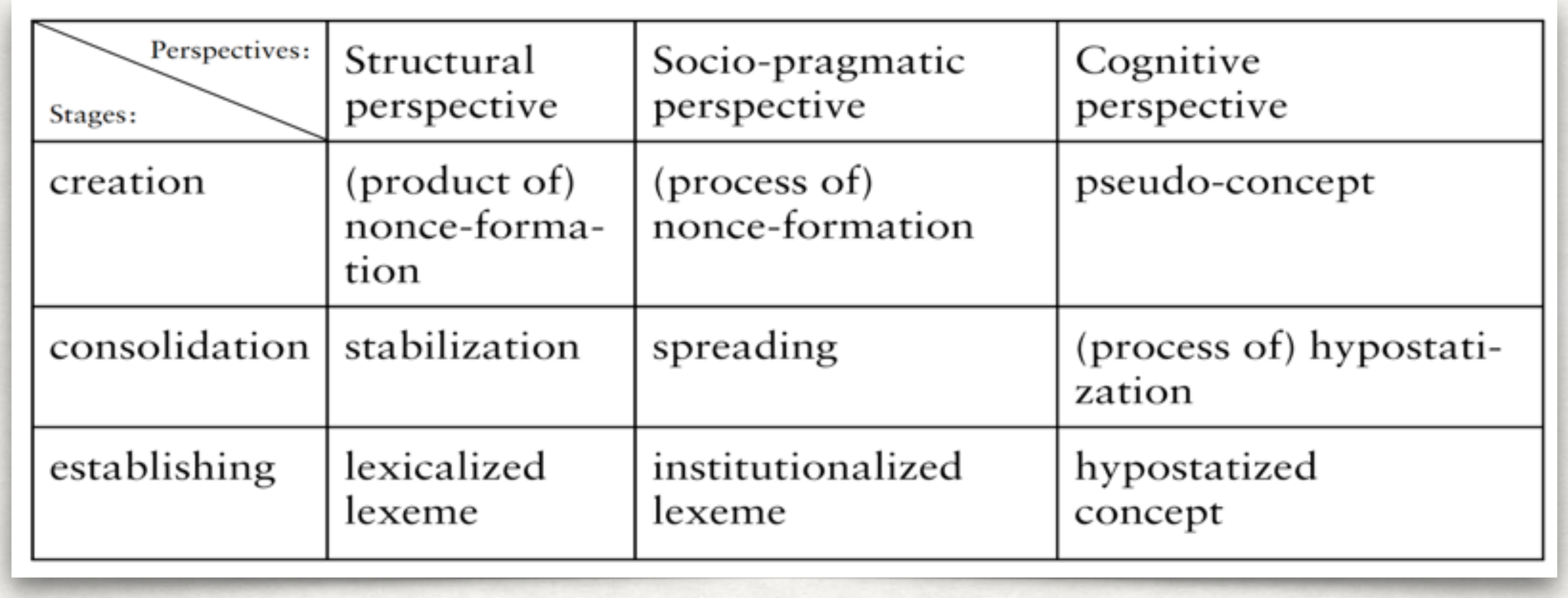
(TOO QUICK) THEORETICAL ASSUMPTIONS : DYNAMICS OF LANGUAGE

(Schmid 2008 ; 2015) three perspectives on lexical innovations :

- **linguistic *perspective :*** *describe the phonological, morphological, syntaxical and semantic features of Lexical Units, and the linguistic mechanisms enabling the modification of any or several of these features;*
- **cognitive *perspective :*** *from the entrenchment (and de-entrenchment) mechanisms, explain how lexical units are processed in the mind (from compositional analysis to routinization). Mainly linked to frequency of exposition to occurrences ;*
- **socio-pragmatic *perspective :*** modelize the pragmatic features of discourse, and the speech communities features where lexical innovations emerge and diffuse.

Three main stages of the life-cycle of lexical innovation : emergence, diffusion and lexicalization, from the linguistic point of view

Schmid 2008 : 3)



Perspectives: Stages:	Structural perspective	Socio-pragmatic perspective	Cognitive perspective
creation	(product of) nonce-formation	(process of) nonce-formation	pseudo-concept
consolidation	stabilization	spreading	(process of) hypostatization
establishing	lexicalized lexeme	institutionalized lexeme	hypostatized concept

NEOVEILLE

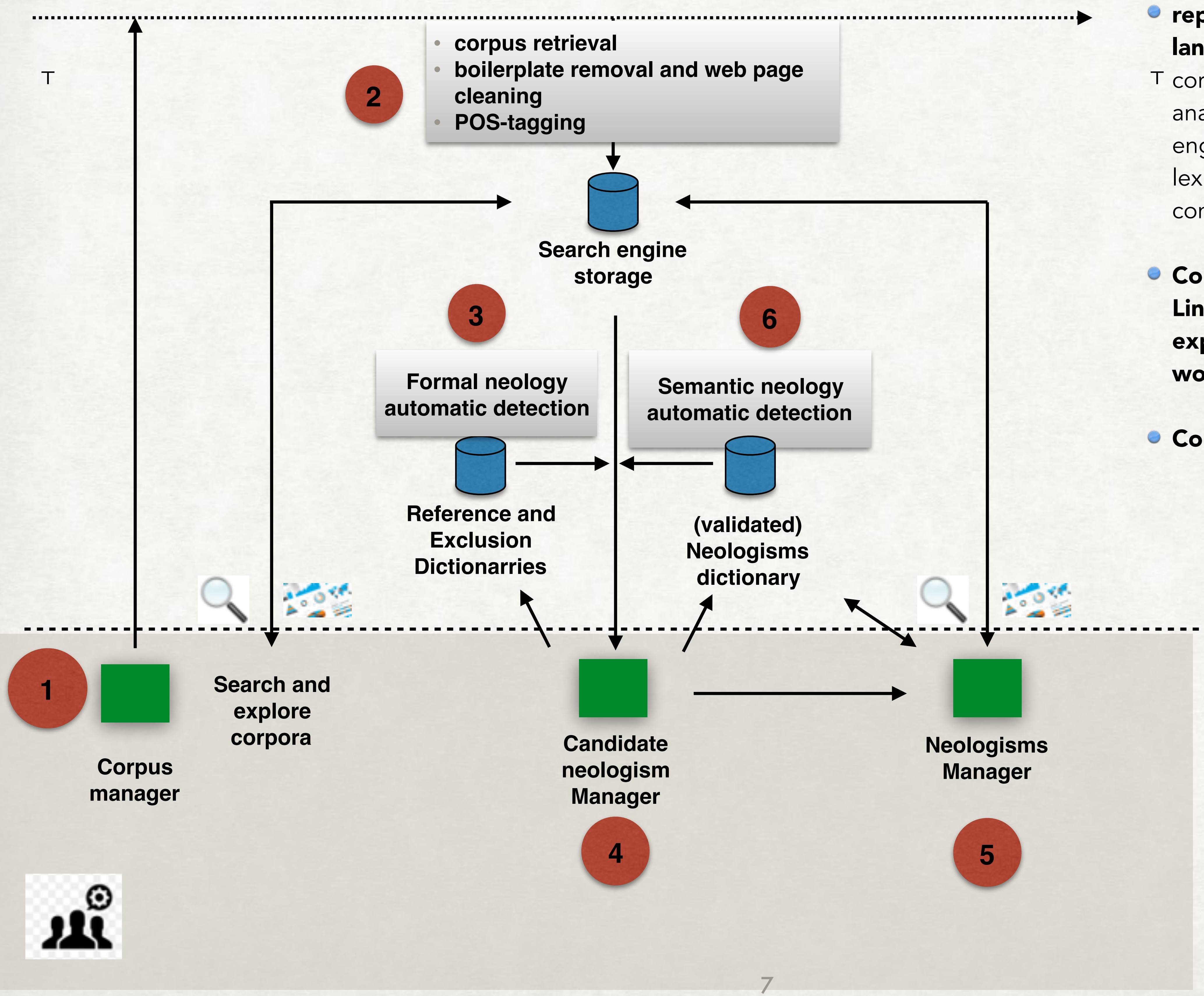
- **Funded Research project 2015-2018 gathering 7 research center (LIPN, CLILLAC-ARP, HTL in France)**

- **main goal : setup a web platform to detect neologisms and track their lifecycle through monitor (contemporary) web corpora**

- **seven languages (French, Brasilian Portuguese, Czech, Greek, Polish, Russian, Chinese), recently extended to Italian, German, Dutch and Spanish**

- **www.neoveille.org, with results freely available on the public part, and a private area for editing and additional features.**

NEOVEILLE : ARCHITECTURE



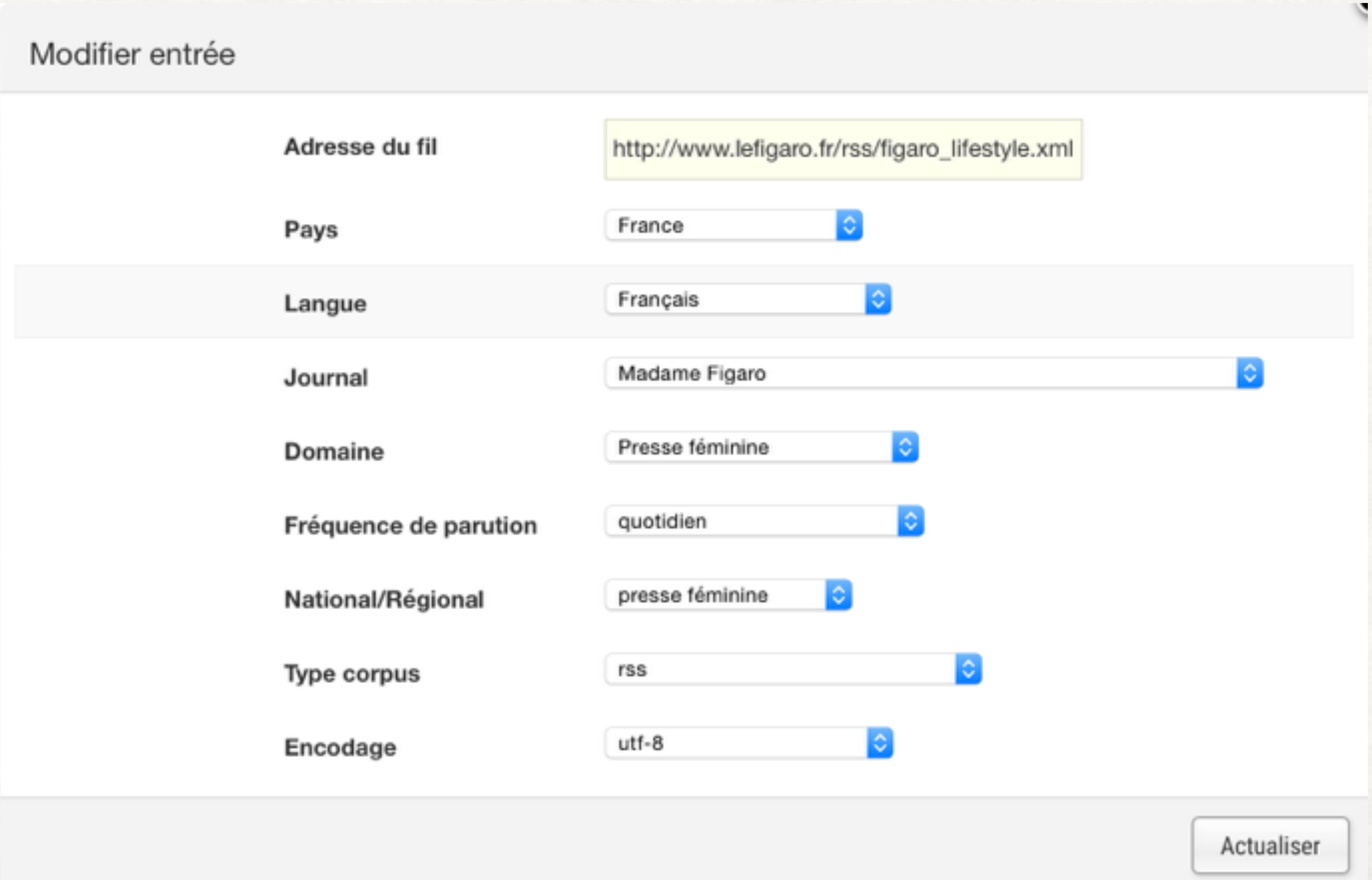
T
2
corpus retrieval
boilerplate removal and web page cleaning
POS-tagging
Search engine storage
3
Formal neology automatic detection
6
Semantic neology automatic detection
Reference and Exclusion Dictionarries
(validated) Neologisms dictionary
1
Corpus manager
Search and explore corpora
Candidate neologism Manager
4
Neologisms Manager
5

- **reproduce the discourse/ language interaction :** monitor corpora from the web, automatic analysis and storage in search engine and databases for lexicographical data (or constructicon!).
- **Combine Computational Linguistics and Human expertise (and give the last word to humans!)**
- **Components :**
  - corpus manager
  - Automatic analysis of articles, storage in search engine, automatic detection of neologisms
  - Manual validation of candidate neologisms
  - Neologisms manager : linguistic description of validated neologisms
  - Visualization Tools to track the lifecycle of neologisms

NEOVEILLE : MODULES - CORPUS MANAGER

Corpus manager :

- basic functionalities : add, read, edit, delete, search
- every source of information has metadata to explicit diastratic (domain), diatopic (region or country)
- at the moment, working with press articles only
- once saved, every source of information is automatically retrieved twice a day, POS-tagged and stored in the search engine.



Modifier entrée

Adresse du fil	http://www.lefigaro.fr/rss/figaro_lifestyle.xml
Pays	France
Langue	Français
Journal	Madame Figaro
Domaine	Presse féminine
Fréquence de parution	quotidien
National/Régional	presse féminine
Type corpus	rss
Encodage	utf-8

NEOVEILLE : MODULES - CORPUS MANAGER

French : 249 RSS feeds

- 154 French (mainland) and 18 local journals
- 50 generalist newspapers, others domain-focused

+1 600 000 articles retrieved in French since sept. 2015

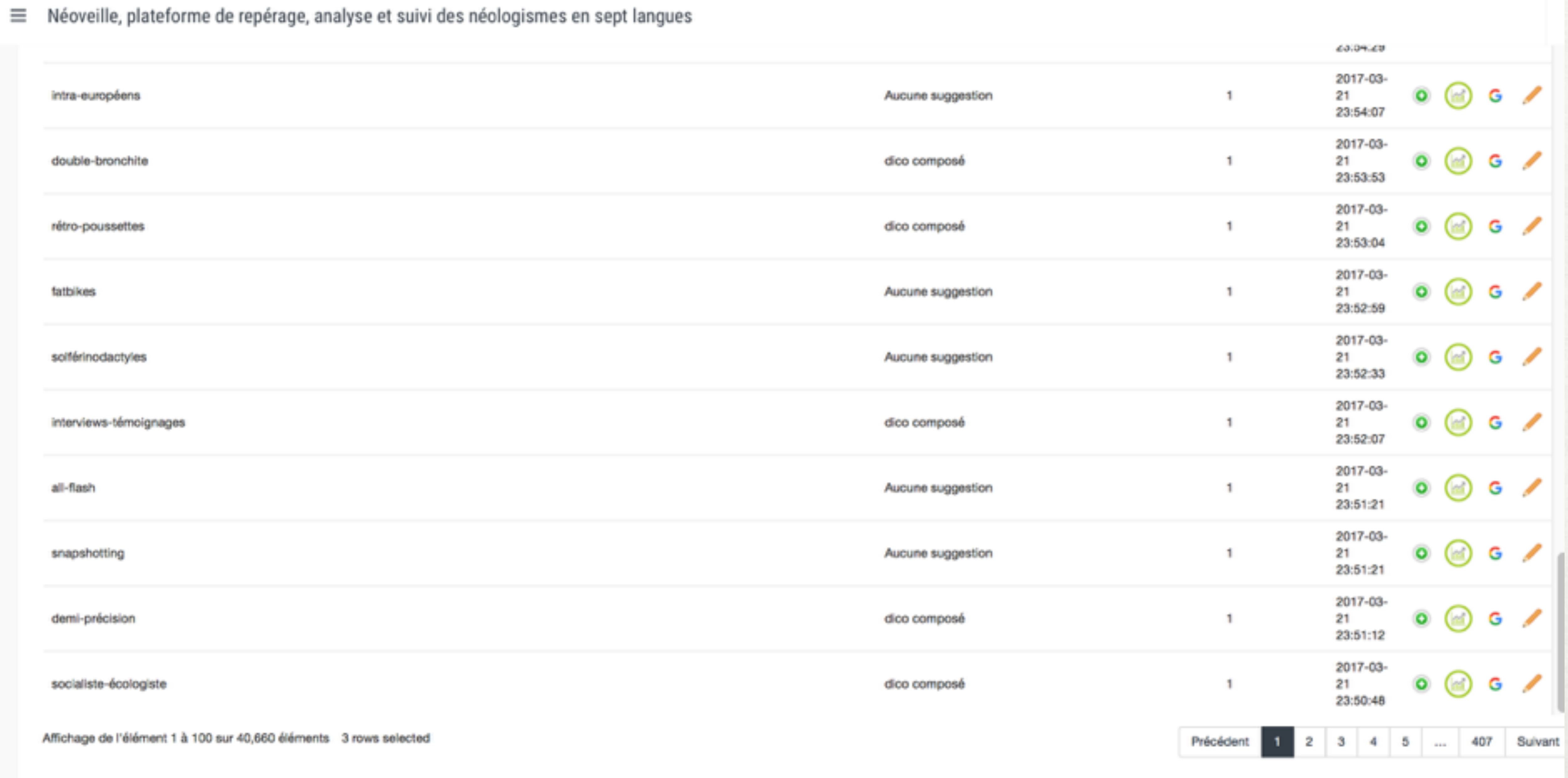
Main domains (from the IIPTC typology): Sports, computing, mode, politics, sciences, Health…

Diatopy from 2016 on with :
Canada, Belgium, Swiss, Algeria, Marocco, Sénégal

NEOVEILLE : MODULES - CANDIDATE-NEOLOGISMS MANAGER

Manual validation of candidate neologisms (CN)

- Formal neology detection : exclusion dictionary method + filters (spelling mistakes, Proper names, citations in foreign languages etc.)
- Néoveille : between 100 and 200 LU per day, among which about 60% are true neologisms
- web interface to validate CN : process enabling to feed the neologism database but also specific exclusion dictionaries (bootstrap process)



≡ Néoveille, plateforme de repérage, analyse et suivi des néologismes en sept langues

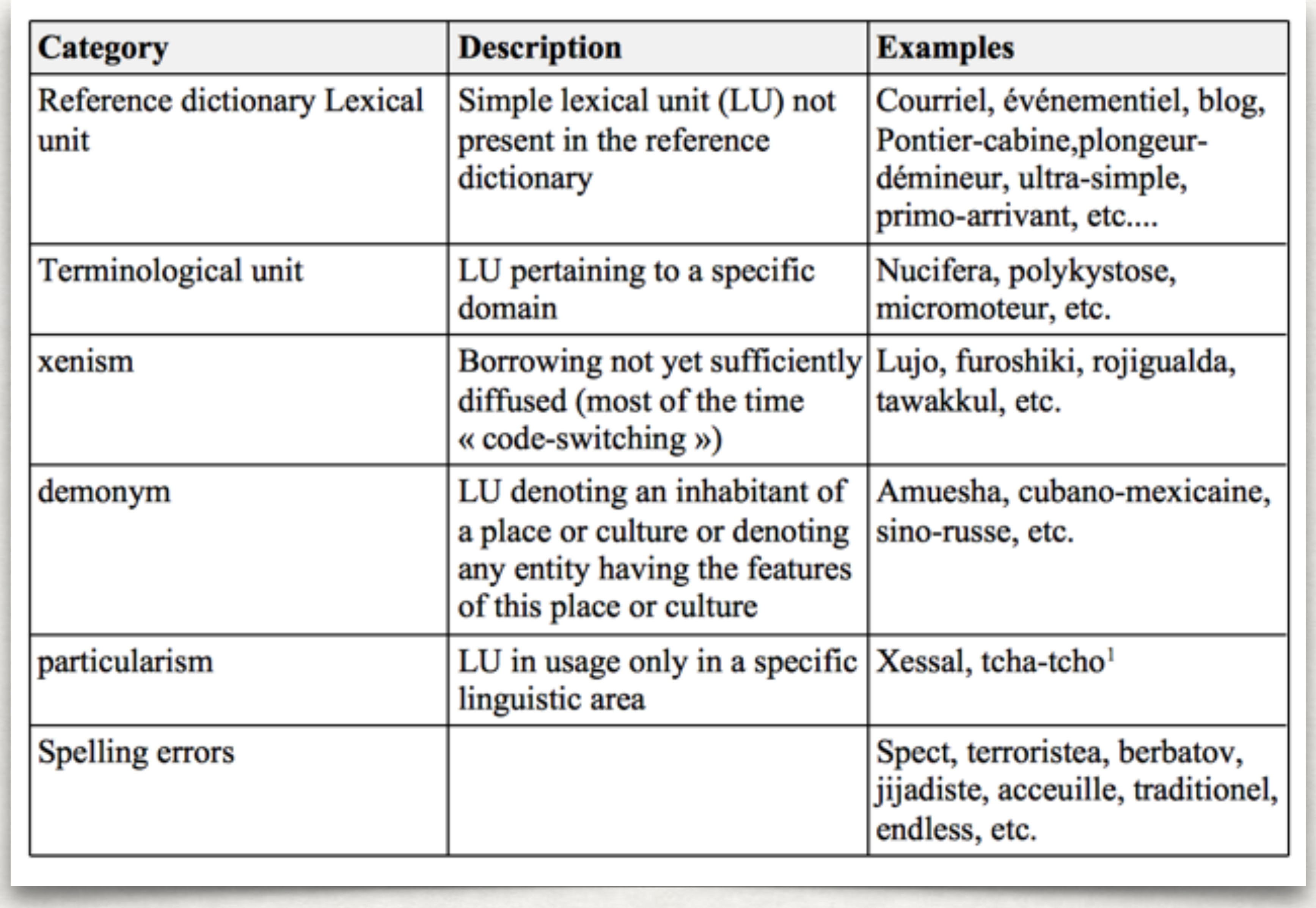
intra-européens	Aucune suggestion	1	2017-03-21 23:54:07
double-bronchite	dico composé	1	2017-03-21 23:53:53
rétro-poussettes	dico composé	1	2017-03-21 23:53:04
fatbikes	Aucune suggestion	1	2017-03-21 23:52:59
solférinodactyles	Aucune suggestion	1	2017-03-21 23:52:33
interviews-témoignages	dico composé	1	2017-03-21 23:52:07
all-flash	Aucune suggestion	1	2017-03-21 23:51:21
snapshotting	Aucune suggestion	1	2017-03-21 23:51:21
demi-précision	dico composé	1	2017-03-21 23:51:12
socialiste-écologiste	dico composé	1	2017-03-21 23:50:48

Affichage de l'élément 1 à 100 sur 40,660 éléments 3 rows selected

NEOVEILLE : MODULES - CANDIDATE-NEOLOGISMS MANAGER

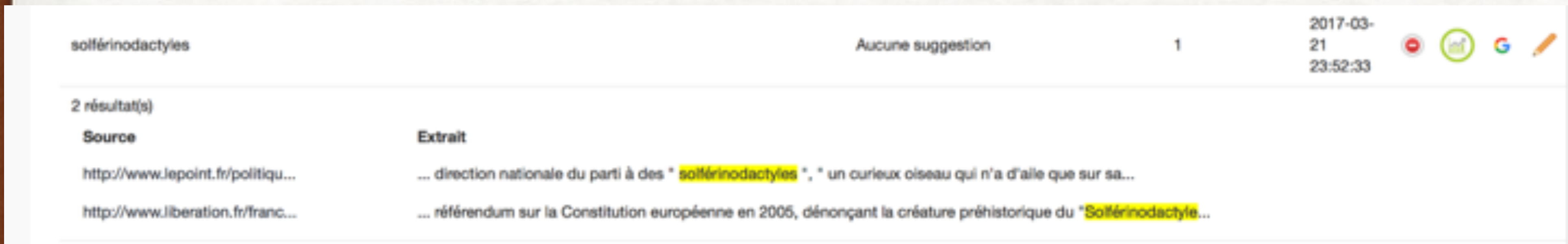
Categorization of candidate neologisms (CN)

- several categories of non-neologisms have been pragmatically designed
- as for neologisms, use of (Sablayrolles, 2000, 2017) typology, refining the tripartition derivation, composition and borrowing
- methodology : group of linguist experts (7 for French) individually annotating with a collective validation



Category	Description	Examples
Reference dictionary Lexical unit	Simple lexical unit (LU) not present in the reference dictionary	Courriel, événementiel, blog, Pontier-cabine,plongeur-démineur, ultra-simple, primo-arrivant, etc....
Terminological unit	LU pertaining to a specific domain	Nucifera, polykystose, micromoteur, etc.
xenism	Borrowing not yet sufficiently diffused (most of the time « code-switching »)	Lujo, furoshiki, rojigualda, tawakkul, etc.
demonym	LU denoting an inhabitant of a place or culture or denoting any entity having the features of this place or culture	Amuesha, cubano-mexicaine, sino-russe, etc.
particularism	LU in usage only in a specific linguistic area	Xessal, tcha-tcho[1]
Spelling errors		Spect, terroristea, berbatov, jijadiste, acceuille, traditionel, endless, etc.

NEOVEILLE : MODULES - CANDIDATE-NEOLOGISMS MANAGER

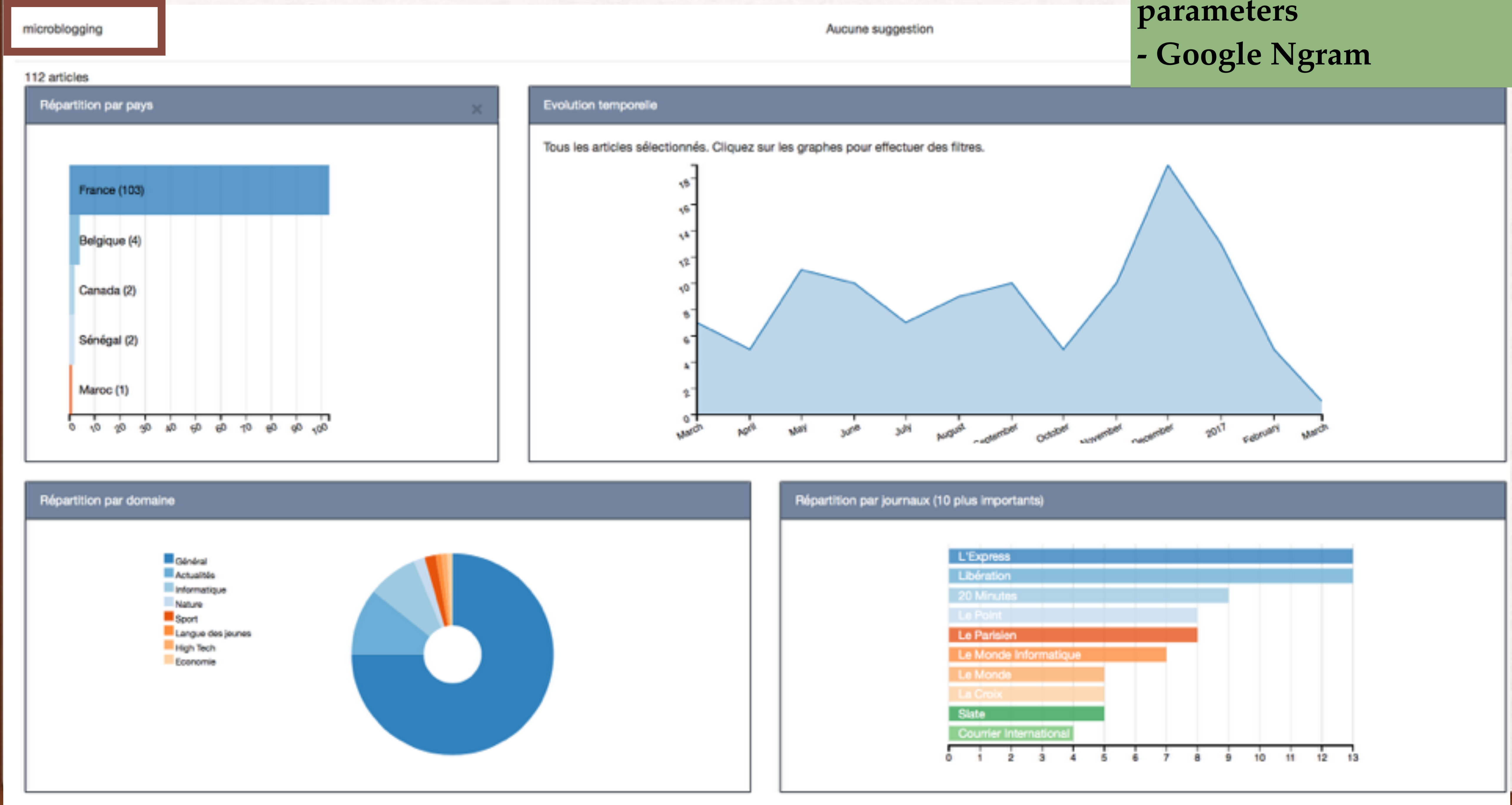


solférinodactyles	Aucune suggestion	1	2017-03-21 23:52:33

2 résultat(s)

Source	Extrait
http://www.lepoint.fr/politiqu...	... direction nationale du parti à des " solférinodactyles ", " un curieux oiseau qui n'a d'aile que sur sa...
http://www.liberation.fr/franc...	... référendum sur la Constitution européenne en 2005, dénonçant la créature préhistorique du "Solférinodactyle...

Several not clear-cut cases =>
- access to contexts
- visualization of diachronic, diastratic and diatopic parameters
- Google Ngram



microblogging

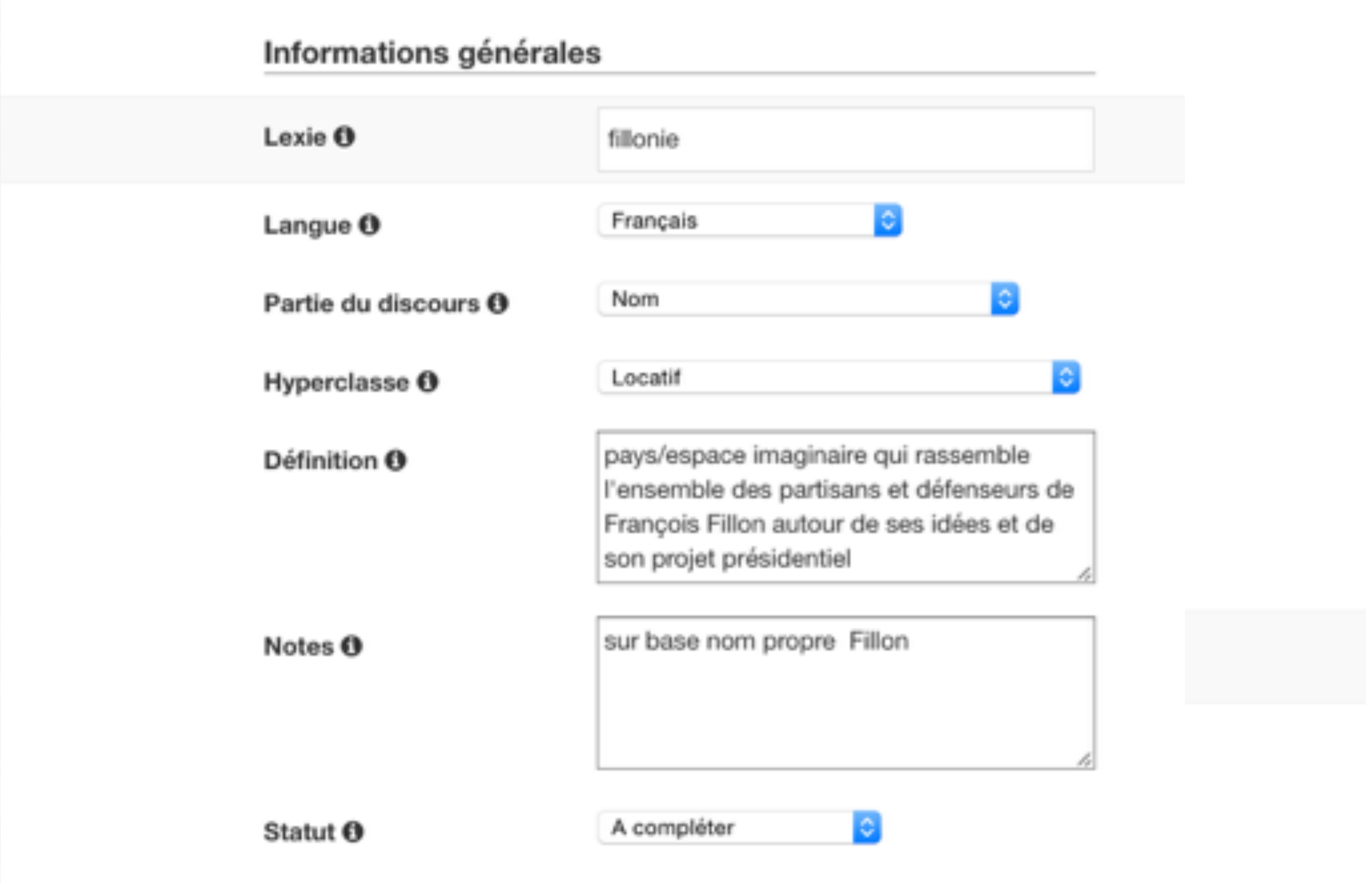
Aucune suggestion

112 articles

NEOVEILLE : MODULES - NEOLOGISMS MANAGER

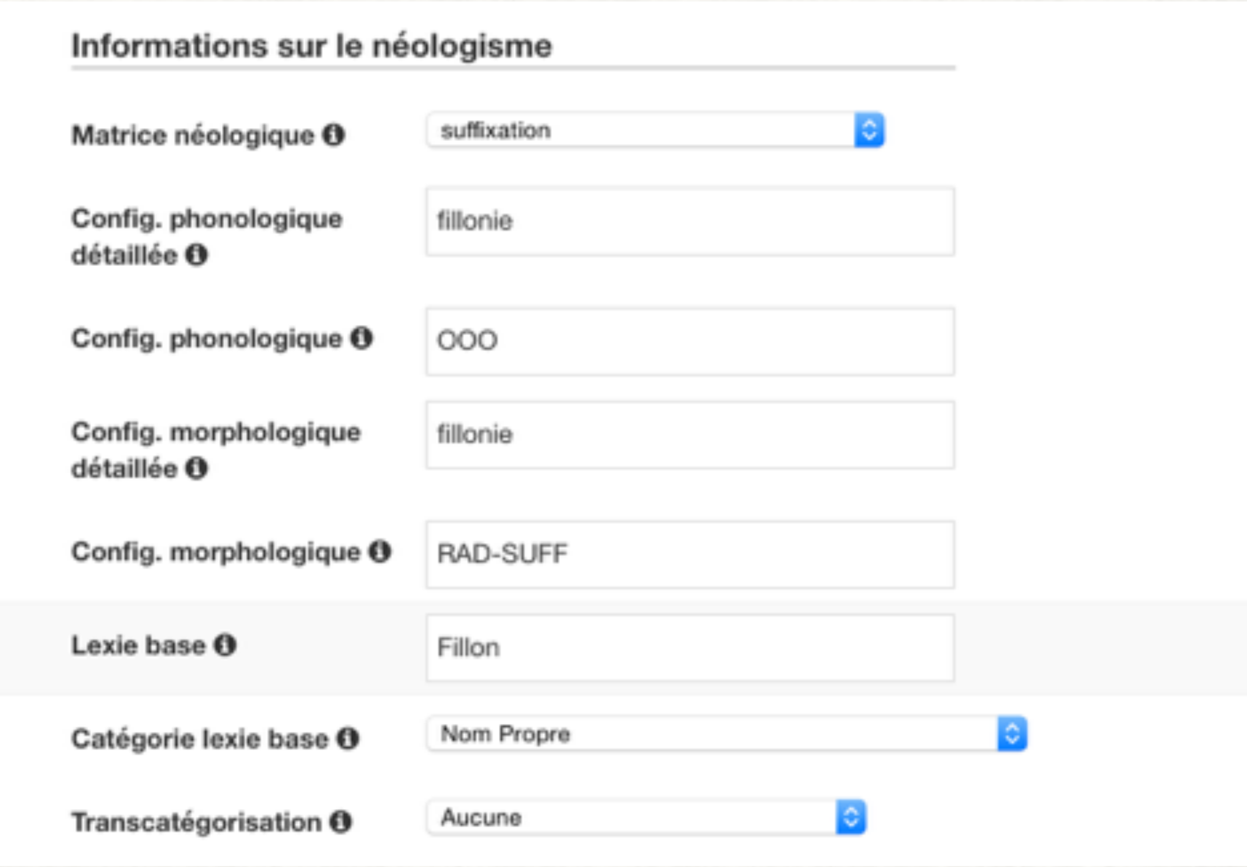
description of validated neologisms

- fine-tuned description (Sablayrolles, 2016; Sablayrolles et Cartier, 2009)
- base of 28 000 neologisms from 2015



Informations générales

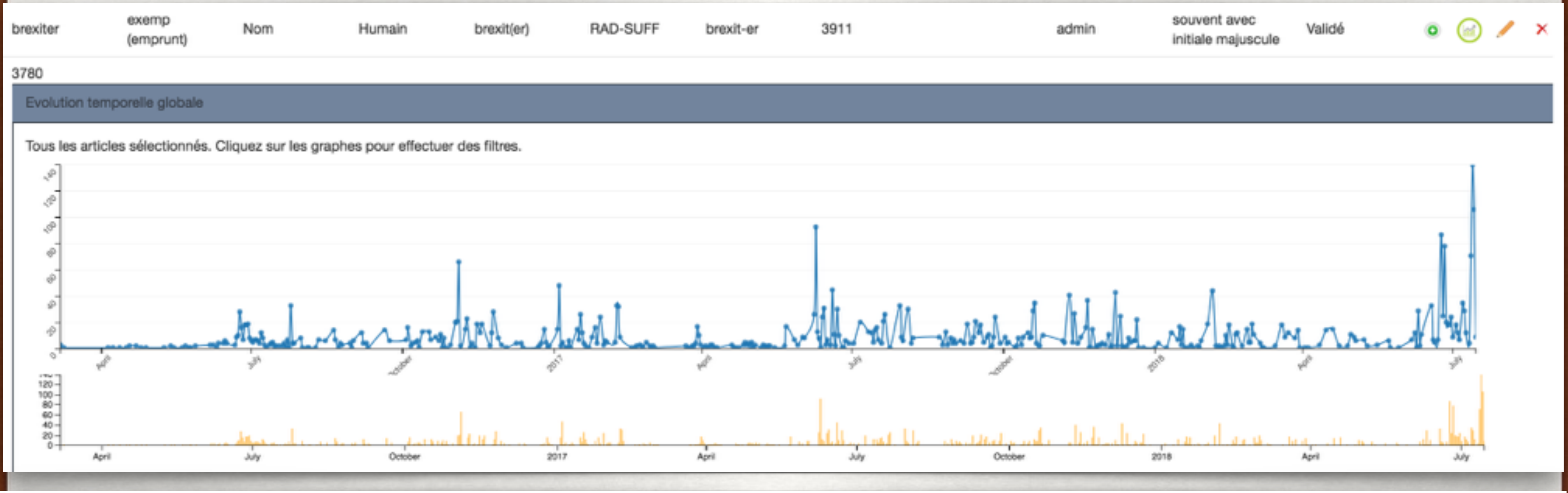
Lexie	fillonie
Langue	Français
Partie du discours	Nom
Hyperclasse	Locatif
Définition	pays/espace imaginaire qui rassemble l'ensemble des partisans et défenseurs de François Fillon autour de ses idées et de son projet présidentiel
Notes	sur base nom propre Fillon
Statut	A compléter



Informations sur le néologisme

Matrice néologique	suffixation
Config. phonologique détaillée	fillonie
Config. phonologique	OOO
Config. morphologique détaillée	fillonie
Config. morphologique	RAD-SUFF
Lexie base	Fillon
Catégorie lexie base	Nom Propre
Transcatégorisation	Aucune

NEOVEILLE : MODULES - LIFE-CYCLE(S) - TEMPORAL AND DIASTRATIC INFORMATION

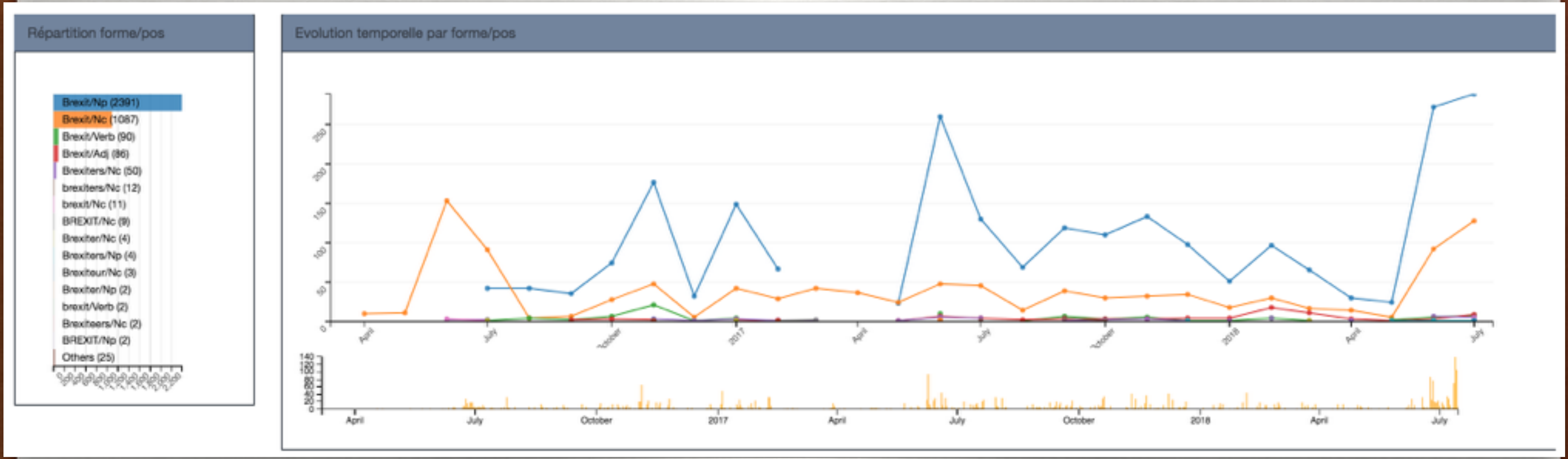


brexiter	exemp (emprunt)	Nom	Humain	brexit(er)	RAD-SUFF	brexit-er	3911	admin	souvent avec initiale majuscule	Validé

3780

Evolution temporelle globale

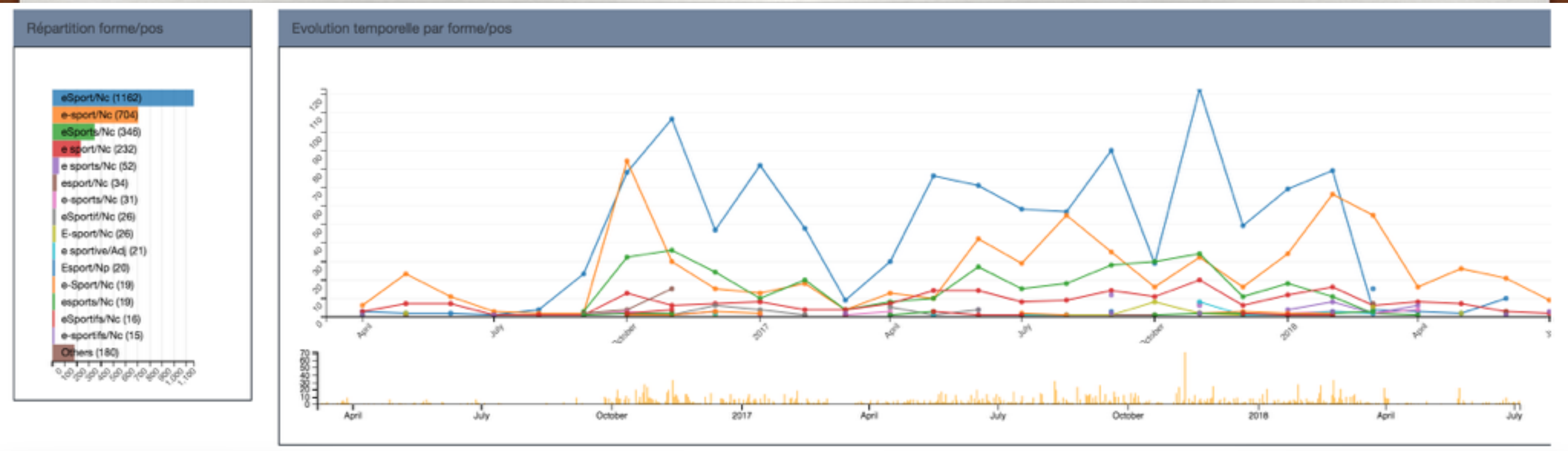
Tous les articles sélectionnés. Cliquez sur les graphes pour effectuer des filtres.



Répartition forme/pos

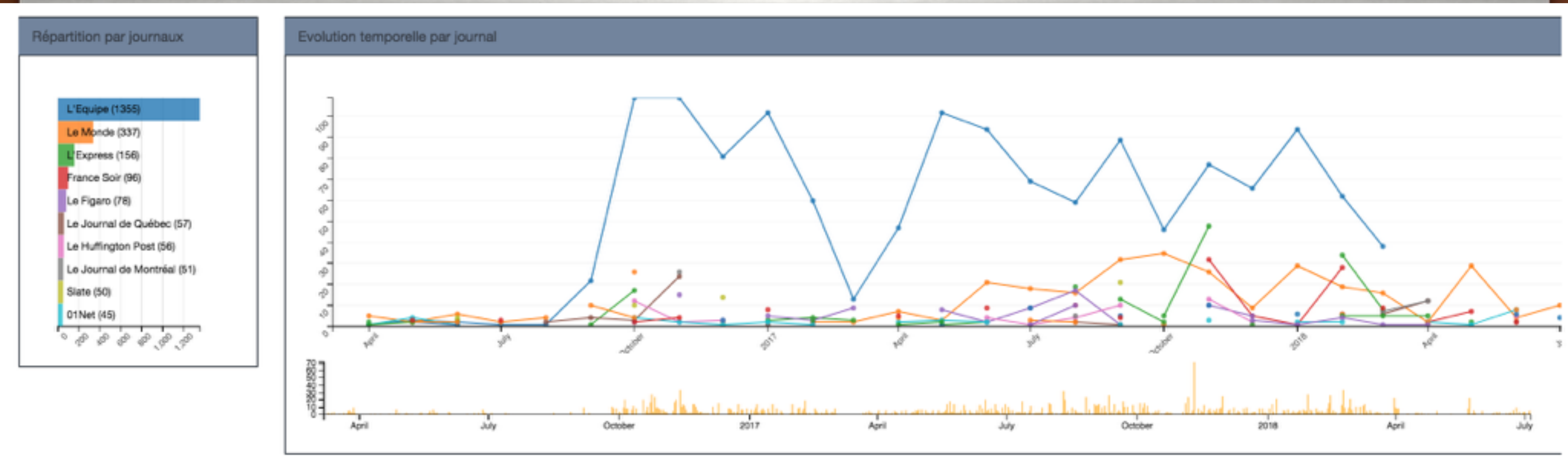
Evolution temporelle par forme/pos

NEOVEILLE : MODULES - LIFE-CYCLE(S) - TEMPORAL AND DIASTRATIC INFORMATION



Répartition forme/pos

Evolution temporelle par forme/pos



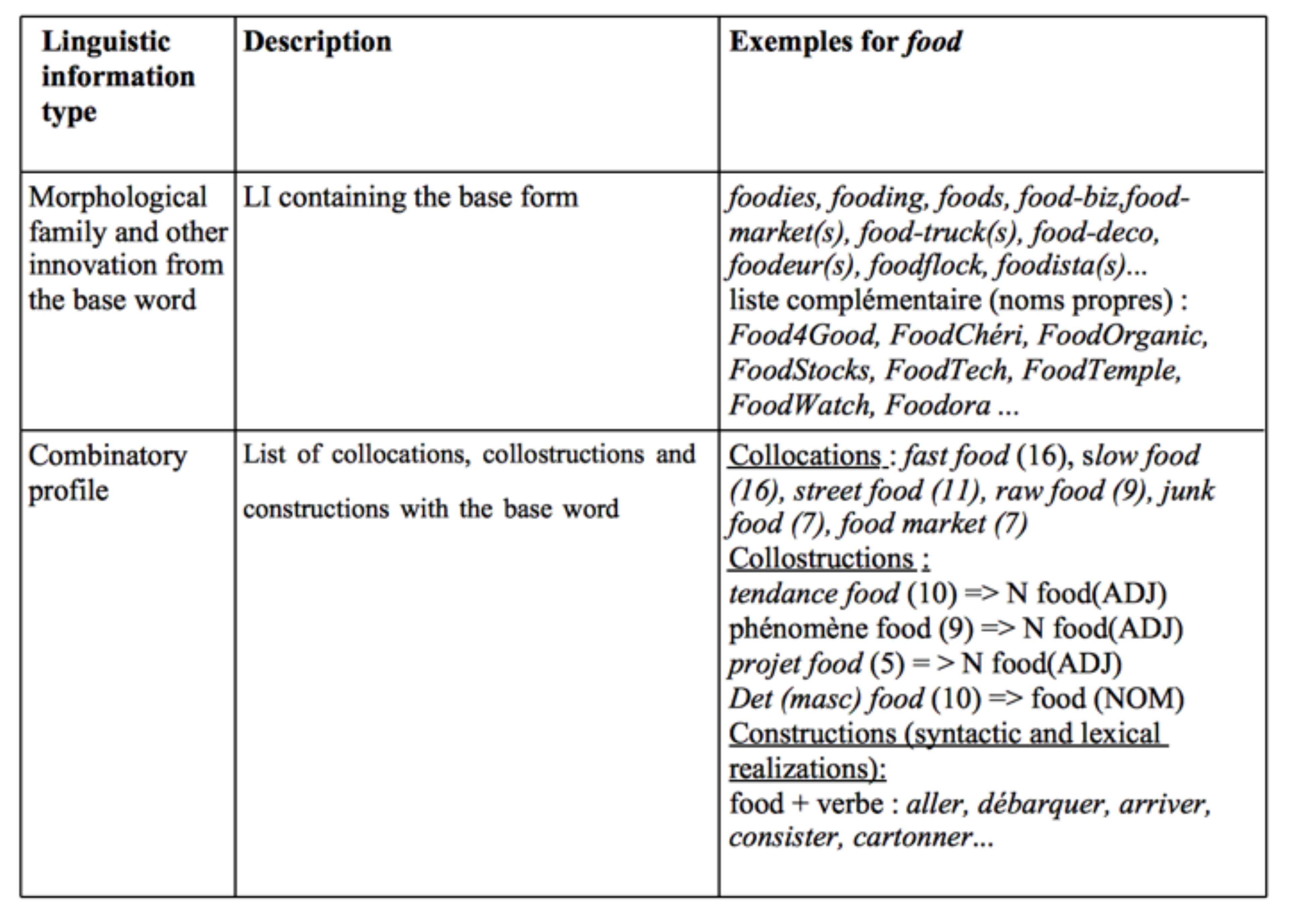
Répartition par journaux

Evolution temporelle par journal

NEOVEILLE : MODULES - LINGUISTIC INFORMATION ON THE LI

Automatic (statistical) linguistic information on the LI

+ distributional semantics profile evolution from the corpus to be added in a near future, enabling to get synonyms, hypernyms and hyponyms and the evolution through time

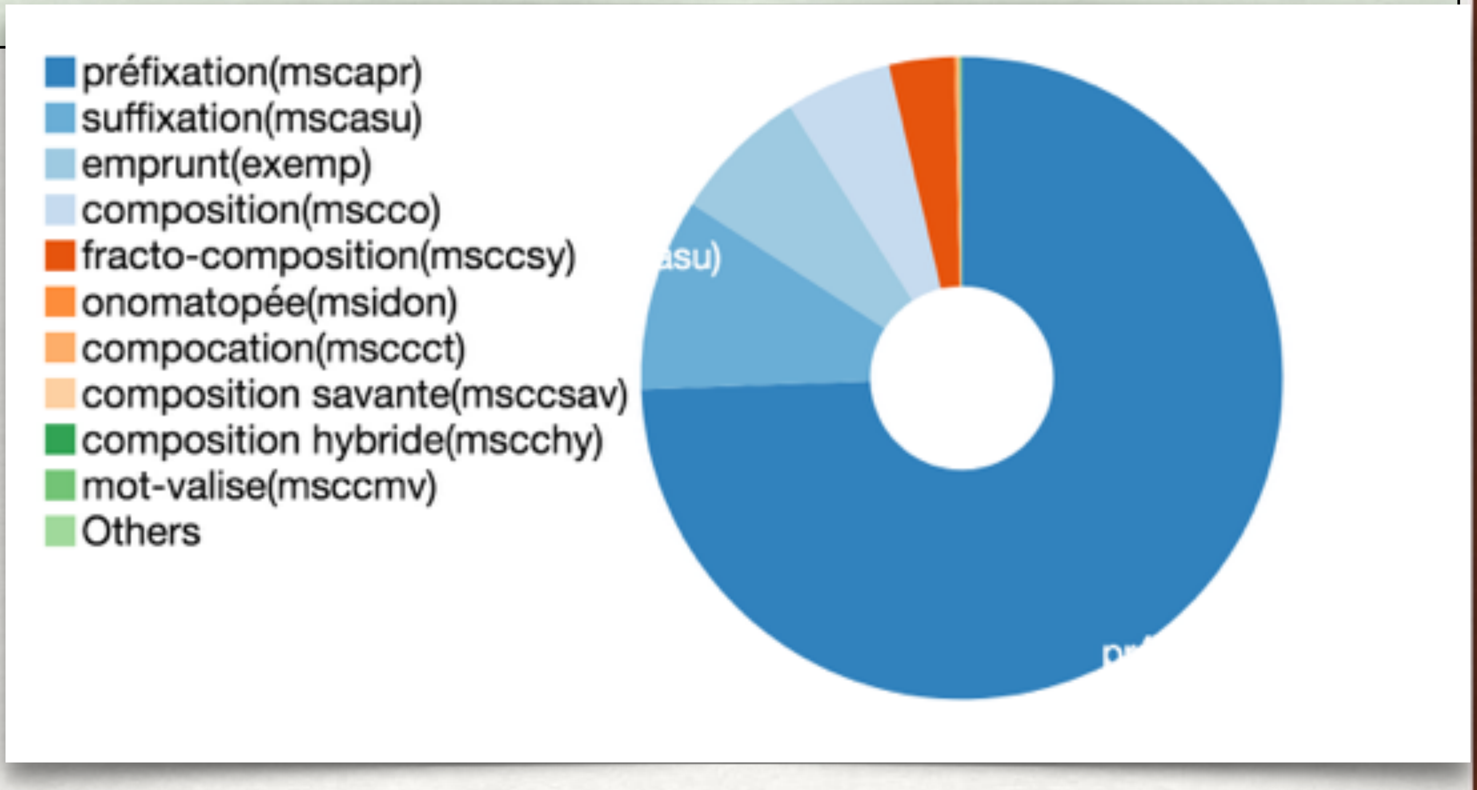


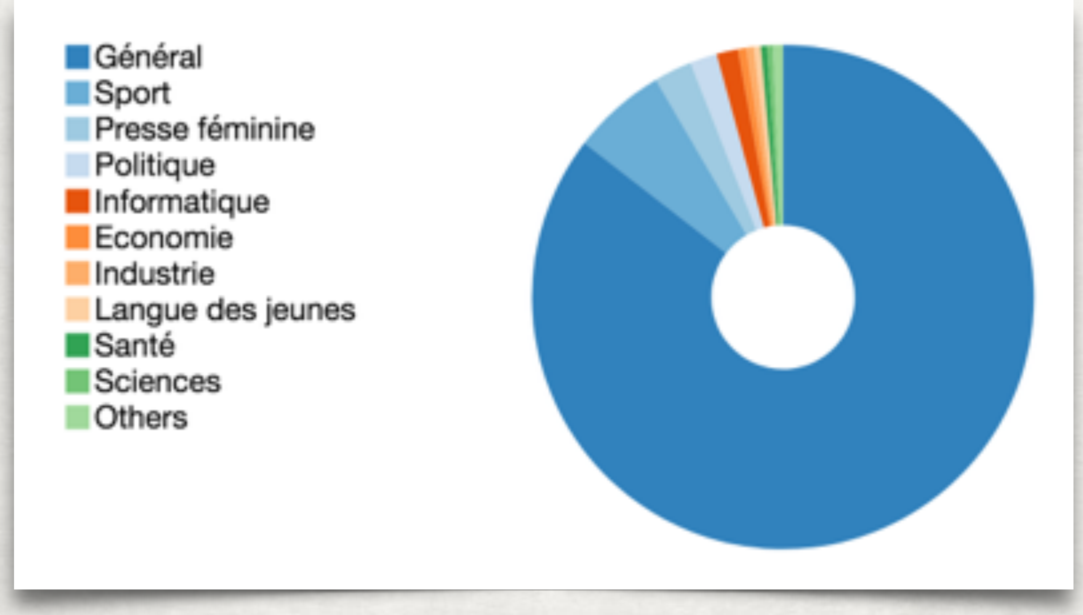
Linguistic information type	**Description**	**Exemples for *food***
Morphological family and other innovation from the base word	LI containing the base form	*foodies, fooding, foods, food-biz,food-market(s), food-truck(s), food-deco, foodeur(s), foodflock, foodista(s)...* liste complémentaire (noms propres) : *Food4Good, FoodChéri, FoodOrganic, FoodStocks, FoodTech, FoodTemple, FoodWatch, Foodora ...*
Combinatory profile	List of collocations, collostructions and constructions with the base word	Collocations : *fast food* (16), *slow food (16), street food (11), raw food (9), junk food (7), food market (7)* Collostructions : *tendance food* (10) => N food(ADJ) phénomène food (9) => N food(ADJ) *projet food* (5) = > N food(ADJ) *Det (masc) food* (10) => food (NOM) Constructions (syntactic and lexical realizations): food + verbe : *aller, débarquer, arriver, consister, cartonner...*

NEOVEILLE : RESULTS FROM FRENCH (2015-2018)

- about 28 000 formal neologisms detected and validated in the period in 250 web sources
- neologisms represent 2,16% (unique word forms) of the whole corpus, and 0,78 % (total number of word forms)
- part-of-speech distribution : Nouns (79,61 %), Adjectives (9,76 %), Verbs (8,34 %) et Adverbs (2,29%).

- prefixation is the most common rule applied (75 %) followed by suffixation (10%), borrowings (8%) and composition (6%)





- Domain distribution enables to identify main innovators
- sports and « feminine press » are mostly using anglo-american borrowings.

NEOVEILLE : RESULTS FROM FRENCH

Hapax and emergence

- normally, we expect that most of neologisms are nonce-words (one occurrence)

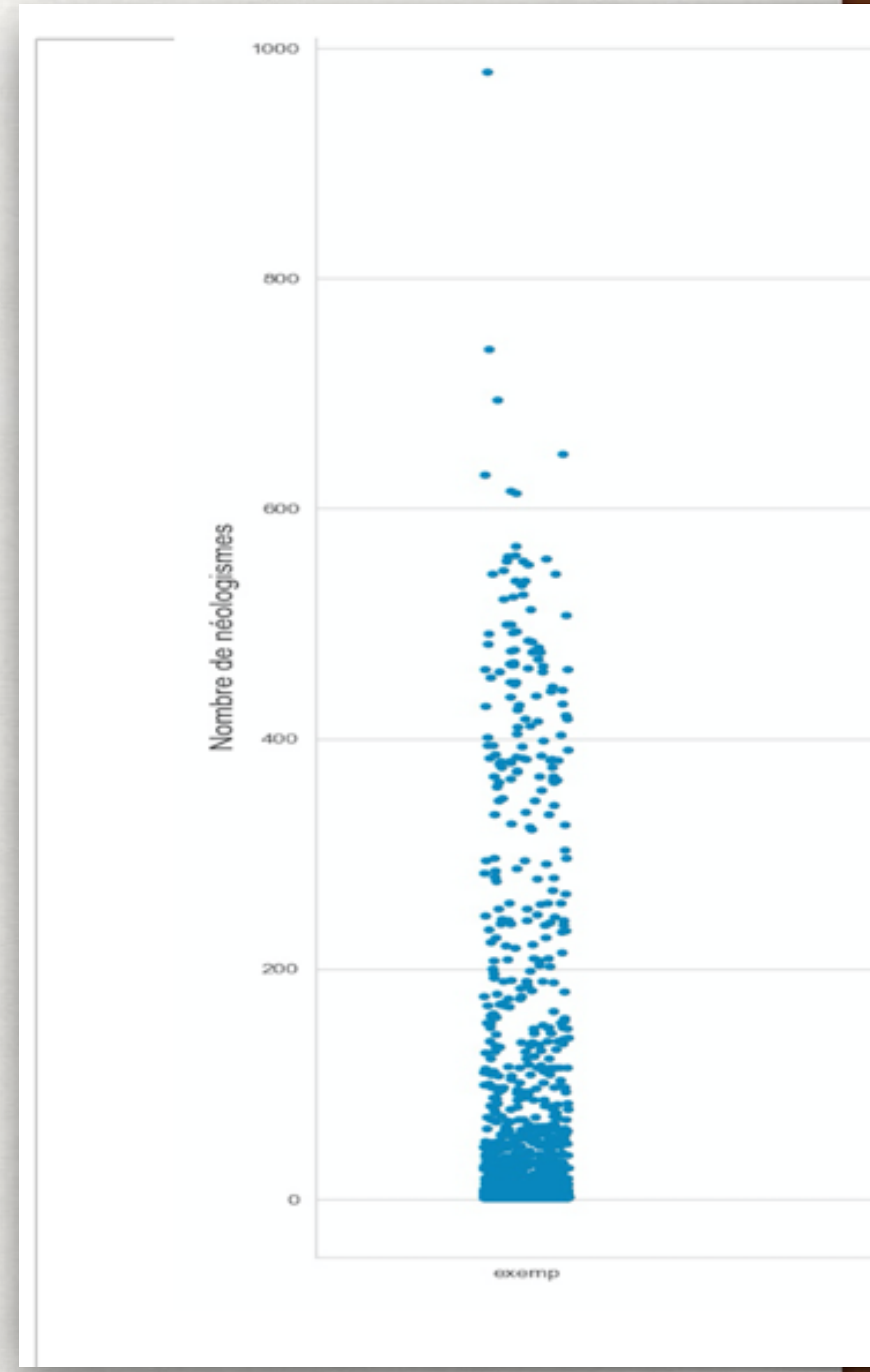
- our data : only 25%, but continuum with the rest of neologisms

- extension of time period to 2 weeks and < 50 occurrences : 70%

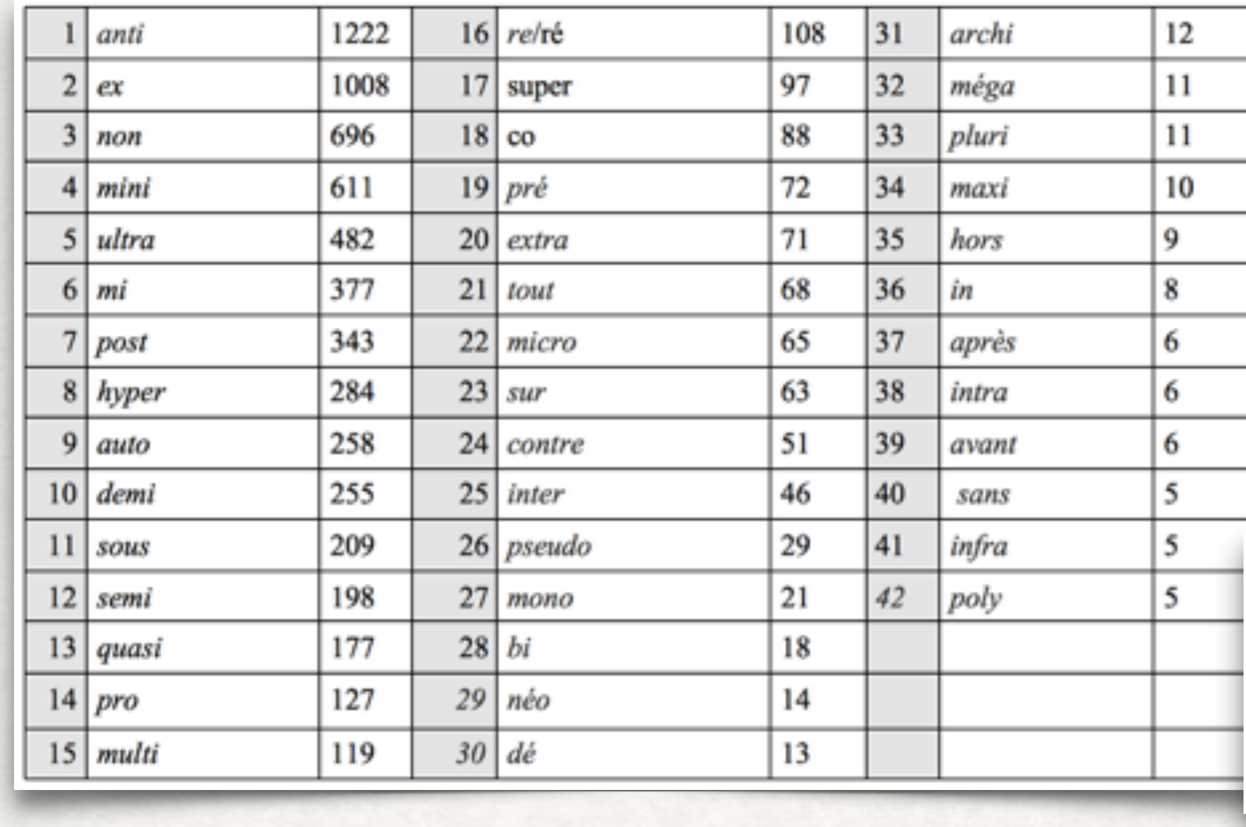
=> most of neologisms are not nonce-words, but neologisms with a very limited diffusion (mainly due to contemporary communication?)

- domains : 70 % are limited to one domain.

=> Extension of domain as a good sign for diffusion



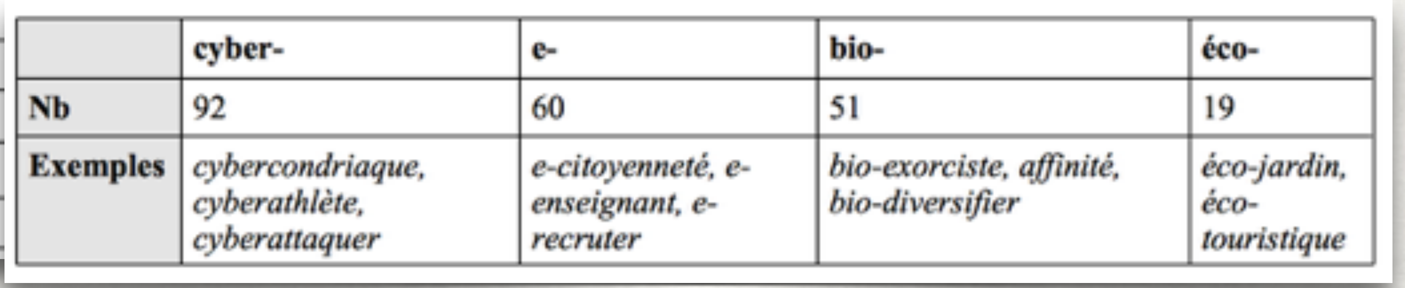
NEOVEILLE : RESULTS FROM FRENCH - PREFIXES, FRACTO-LEXEMES



1	anti	1222	16	re/ré	108	31	archi	12
2	ex	1008	17	super	97	32	méga	11
3	non	696	18	co	88	33	pluri	11
4	mini	611	19	pré	72	34	maxi	10
5	ultra	482	20	extra	71	35	hors	9
6	mi	377	21	tout	68	36	in	8
7	post	343	22	micro	65	37	après	6
8	hyper	284	23	sur	63	38	intra	6
9	auto	258	24	contre	51	39	avant	6
10	demi	255	25	inter	46	40	sans	5
11	sous	209	26	pseudo	29	41	infra	5
12	semi	198	27	mono	21	42	poly	5
13	quasi	177	28	bi	18			
14	pro	127	29	néo	14			
15	multi	119	30	dé	13			

Productivity (Baayen, 2009) :

- **realized *productivity* :** *attested occurrences*
- ***expanding* productivity :** quantification of the newly coined neologisms from the element (here affixes).
- **potential productivity : measure of the maximum capacity of the element to generate new words**, ie depending on the *rule constraints*. (example : *non-* has a greater potential than *ex-* as the first can be applied to nouns and adjectives, whereas the second is limited to nouns)



	cyber-	e-	bio-	éco-
Nb	92	60	51	19
Exemples	cybercondriaque, cyberathlète, cyberattaquer	e-citoyenneté, e-enseignant, e-recruter	bio-exorciste, affinité, bio-diversifier	éco-jardin, éco-touristique

- **top productive (expanding productivity) affixes are those whose potential productivity (in terms of applicable POS and meaning) is the less constrained (anti-, ex-, non-, mini, ultra-, mi-)**
- **verbs less productive (post-,hyper-,auto-, etc.)**
- **emergence of fracto-lexemes in the last 20 years with a quasi-prefix functioning**

NEOVEILLE : RESULTS FROM FRENCH - BORROWINGS

- 1 430 formes (6,36% du total) pour 132 104 occurrences (18,19%) + environ 1 000 xénismes
- langue source la plus représentée est l'**anglo-américain international** (91%), suivi de l'espagnol, de l'arabe et de l'italien. Les xénismes ont des langues–sources beaucoup plus diversifiées.
- Trois **domaines innovateurs** sont particulièrement productifs : presse féminine (*Elle, Grazia, Cosmo, Styles*), informatique (*01Net, Le monde informatique*) et sport (*L'équipe*).
- **Les emprunts à l'anglais ne se limitent plus au transfert de lexies** :
  - du point de vue phonologique et orthographique, l'influence de l'anglo-américain est perceptible depuis longtemps (prononciation de *-ing*, *-ee-*, etc.).
  - **implantation d'affixes**, notamment le fracto-lexème *e-* et le suffixe *-ing*. (e- : 86 lexies pour le premier (soit emprunts directs : *e-voting*, *e-shopping*, etc. soit hybrides : *e-défilé*, *e-vendeur*, *e-marché*, *e-citoyenneté*, etc.); -ing : 303. concurrence avec *-age* fait qu'il reste limité à l'expression de pratiques sportives (*running*, *beatboxing*, *snorkeling*, *cardiotraining*, etc.) professionnelles (*networking*, *packaging*, *branding*, *fact-checking*, *coworking*, *crowdfunding*,…) ou socio-culturelles (*bashing*, *ghosting*, *pet-sitting*) spécifiques, sans équivalents synthétiques en français.
- **émergence de formations et de patrons lexico-syntaxiques productifs** : formations en ***-gate*** (56 occurrences : *dieselgate*, *couscousgate*, *penaltygate*, *penelopegate*, etc.) ; ***street-*** (25 lexies: *streetstyle*, *street-artiste*, etc.) ; ***food-*** (23 lexies : *food-truck*, *foodosphère*, *foodocratie*, *foodivores*, *street-fooders*, etc.) ; ***-bashing*** (11 lexies: *agribashing*, *sucre-bashing*, *macronbashing*, *etc.*), ***-shaming*** (14 lexies: *fatshaming*, *name-shaming*, *skillshaming*), ***it-*** (8 lexies : *it-jean*, *it-bag*, etc.). Nous relevons également 144 occurrences du patron **N/ADJ-V*ing*** (*car-jacking*, *home-staging*, *speed-dating*, *speed-watching*, *binge-viewing*, *ride-sharing*).

CONCLUSION

Web platform currently operational : www.neoveille.org

Corpora have been setup for 11 languages and documents are retrieved daily, with rich linguistic and metainformation linked to source documents

Active groups of researchers working on French, Italian and Russian for neologism detection

Several modules :

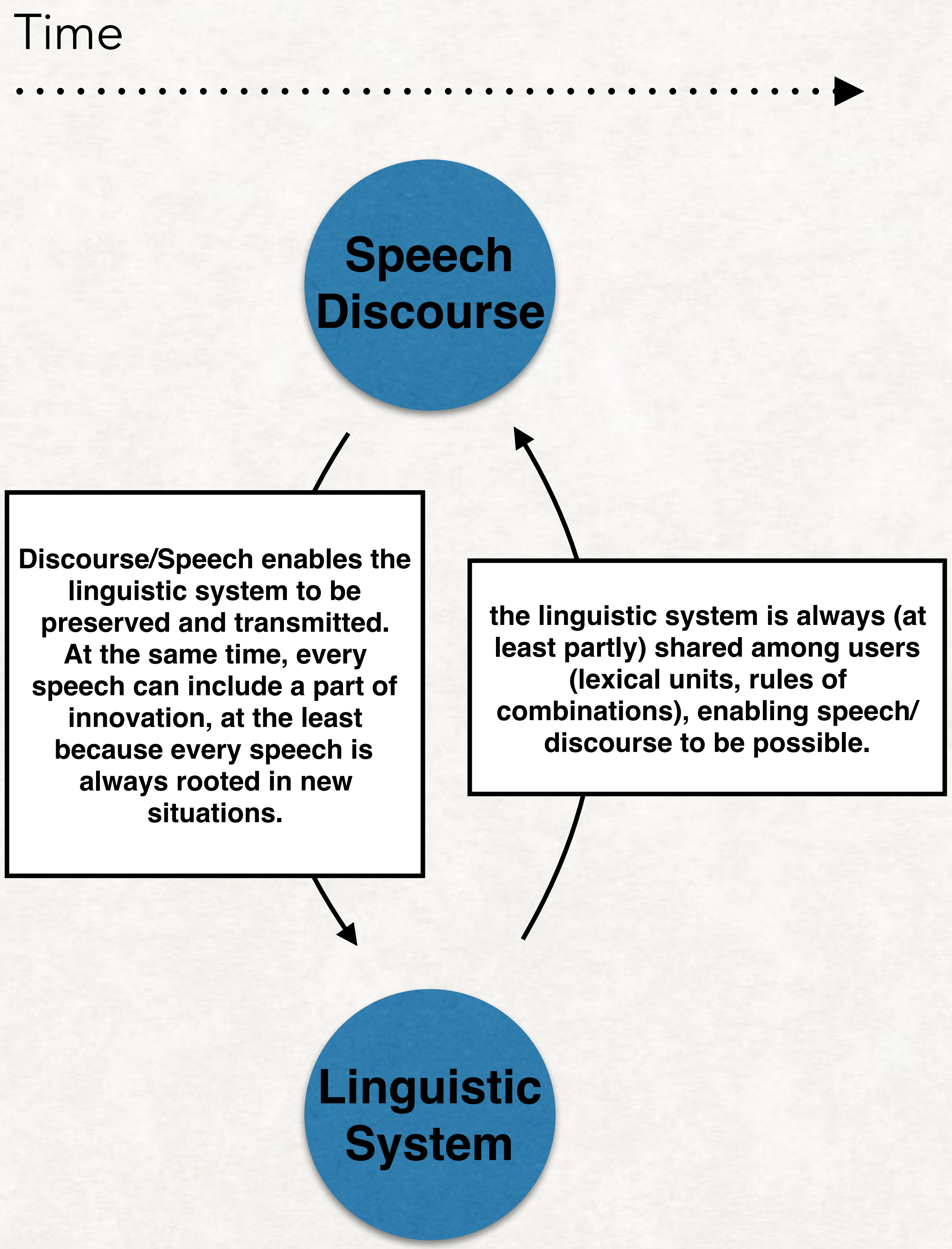
- **corpus manager**
- **neologism automatic detection and validation in a bootstrap process**
- **linguistic, socio-pragmatic and temporal information enable to follow the life of lexical units**
- **methodology enabling to validate candidate neologisms from corpora AND update existing reference dictionary**

Improvements on the track

- **linguistic information : distributional semantics**
- **socio-pragmatic features of source documents need refinement**
- **domains are hard-coded by linguists (from the press editors information) and automatic topic detection can help refine the information**
- **extension of web corpus to blogs and other types of texts**

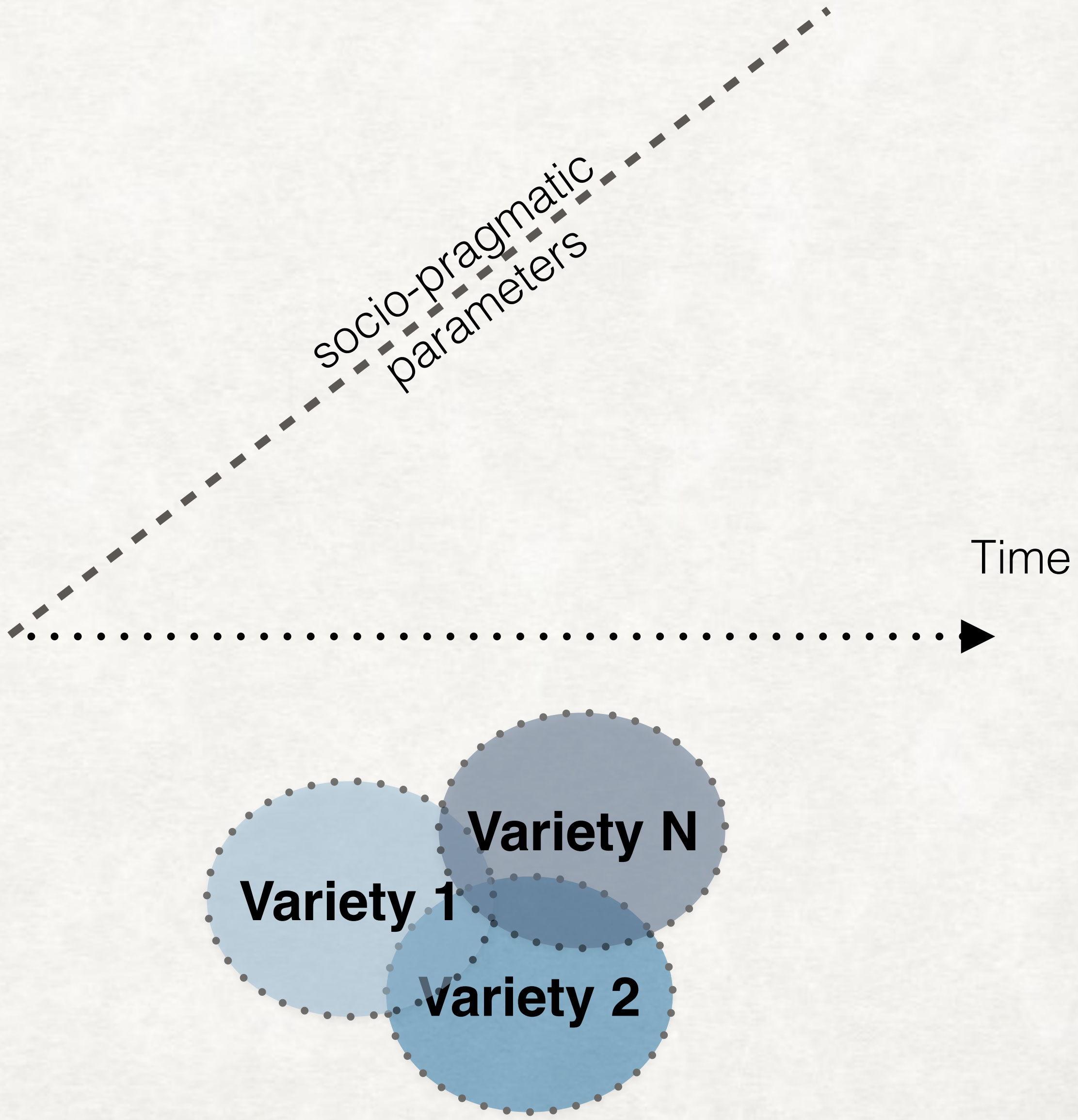
Web platform will soon be open sourced so that research groups can use it for their own research.

(TOO QUICK) THEORETICAL ASSUMPTIONS : DYNAMICS OF LANGUAGE



Discourse/Speech enables the linguistic system to be preserved and transmitted. At the same time, every speech can include a part of innovation, at the least because every speech is always rooted in new situations.

the linguistic system is always (at least partly) shared among users (lexical units, rules of combinations), enabling speech/ discourse to be possible.


socio-pragmatic
parameters
Time
Variety N
Variety 1
Variety 2