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

* **linguistic perspective** : describe the phonological, morphological, syntactical 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)

<table>
<thead>
<tr>
<th>Perspectives:</th>
<th>Structural perspective</th>
<th>Socio-pragmatic perspective</th>
<th>Cognitive perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stages:</td>
<td>creation</td>
<td>consolidation</td>
<td>establishing</td>
</tr>
<tr>
<td></td>
<td>(product of) nonce-formation</td>
<td>stabilization</td>
<td>lexicalized lexeme</td>
</tr>
<tr>
<td></td>
<td>(process of) nonce-formation</td>
<td>spreading</td>
<td>institutionalized lexeme</td>
</tr>
<tr>
<td></td>
<td>pseudo-concept</td>
<td>(process of) hypostatization</td>
<td>hypostatized concept</td>
</tr>
</tbody>
</table>
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, Brasilians 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.
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
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.
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
**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)
**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

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference dictionary</td>
<td>Simple lexical unit (LU) not present in the reference dictionary</td>
<td>Courriel, événementiel, blog, Pontier-cabine, plongeur-démeneur, ultra-simple, primo-arrivant, etc.</td>
</tr>
<tr>
<td>Lexical unit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terminological unit</td>
<td>LU pertaining to a specific domain</td>
<td>Nucifera, polykystose, micromotueur, etc.</td>
</tr>
<tr>
<td>xenism</td>
<td>Borrowing not yet sufficiently diffused (most of the time « code-switching »)</td>
<td>Lujo, furoshiki, rojigualda, tawakkul, etc.</td>
</tr>
<tr>
<td>demonym</td>
<td>LU denoting an inhabitant of a place or culture or denoting any entity having the features of this place or culture</td>
<td>Amuesha, cubano-mexicaine, sino-russe, etc.</td>
</tr>
<tr>
<td>particularism</td>
<td>LU in usage only in a specific linguistic area</td>
<td>Xessal, tcha-tcho¹</td>
</tr>
<tr>
<td>Spelling errors</td>
<td></td>
<td>Spect, terroristea, berbatov, jijadiste, acceuille, traditionel, endless, etc.</td>
</tr>
</tbody>
</table>
Several not clear-cut cases =>
- access to contexts
- visualization of diachronic, diastratic and diatopic parameters
- Google Ngram
description of validated neologisms
- fine-tuned description (Sablayrolles, 2016; Sablayrolles et Cartier, 2009)
- base of 28 000 neologisms from 2015
Evolution temporelle globale

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

Répartition forme/pos

Evolution temporelle par forme/pos

| Brexit/Np (238) | Brex/Np (1087) | Brex/Verb (90) | Brex/Ad (86) | Brex/Nc (50) | Brex/Np (12) | brex/Nc (11) | BREXIT/Nc (9) | Brexer/Nc (4) | Brexera/Np (4) | Brexheur/Nc (3) | Brexer/Np (2) | brex/Verb (2) | Brexeha/Nc (2) | BREXIT/Np (2) | Others (25) |
### NEOVILLE: MODULES - LIFE-CYCLE(S) - TEMPORAL AND DIASTRATIC INFORMATION

#### Répartition forme/pos

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>eSport/Nc</td>
<td>1160</td>
</tr>
<tr>
<td>'e-sport/Nc'</td>
<td>794</td>
</tr>
<tr>
<td>eSports/Nc</td>
<td>346</td>
</tr>
<tr>
<td>e-sport/Nc</td>
<td>232</td>
</tr>
<tr>
<td>e-sports/Nc</td>
<td>52</td>
</tr>
<tr>
<td>eSports/Adn</td>
<td>24</td>
</tr>
<tr>
<td>e-sports/Adn</td>
<td>31</td>
</tr>
<tr>
<td>eSports/AD</td>
<td>26</td>
</tr>
<tr>
<td>e-sports/AD</td>
<td>26</td>
</tr>
<tr>
<td>e-sports/Ad</td>
<td>21</td>
</tr>
<tr>
<td>Esport/Nc</td>
<td>20</td>
</tr>
<tr>
<td>e-sport/Nc</td>
<td>19</td>
</tr>
<tr>
<td>eSports/Nc</td>
<td>19</td>
</tr>
<tr>
<td>e-sports/Nc</td>
<td>16</td>
</tr>
<tr>
<td>e-sports/No</td>
<td>15</td>
</tr>
<tr>
<td>Others</td>
<td>140</td>
</tr>
</tbody>
</table>

#### Evolution temporelle par forme/pos

![Graph showing temporal evolution](image)

#### Répartition par journaux

<table>
<thead>
<tr>
<th>Journal</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>L'Equipe</td>
<td>1355</td>
</tr>
<tr>
<td>Le Monde</td>
<td>337</td>
</tr>
<tr>
<td>L'Express</td>
<td>156</td>
</tr>
<tr>
<td>France Soir</td>
<td>96</td>
</tr>
<tr>
<td>Le Figaro</td>
<td>76</td>
</tr>
<tr>
<td>Le Journal de Québec</td>
<td>57</td>
</tr>
<tr>
<td>Le Huffington Post</td>
<td>54</td>
</tr>
<tr>
<td>Le Journal de Montréal</td>
<td>51</td>
</tr>
<tr>
<td>Slate</td>
<td>50</td>
</tr>
<tr>
<td>Dessin</td>
<td>45</td>
</tr>
</tbody>
</table>

#### Evolution temporelle par journal

![Graph showing temporal evolution](image)
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

<table>
<thead>
<tr>
<th>Linguistic information type</th>
<th>Description</th>
<th>Exemples for food</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological family and other innovation from the base word</td>
<td>LI containing the base form</td>
<td><strong>foodies</strong>, <strong>fooding</strong>, <strong>foods</strong>, <strong>food-biz</strong>, <strong>food-market(s)</strong>, <strong>food-truck(s)</strong>, <strong>food-deco</strong>, <strong>foodeur(s)</strong>, <strong>fooflock</strong>, <strong>foodista(s)</strong>... liste complémentaire (noms propres) : <strong>Food4Good</strong>, <strong>FoodChéri</strong>, <strong>FoodOrganic</strong>, <strong>FoodStocks</strong>, <strong>FoodTech</strong>, <strong>FoodTemple</strong>, <strong>FoodWatch</strong>, <strong>Foodora</strong>...</td>
</tr>
</tbody>
</table>
| Combinatory profile | List of collocations, collocations 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)
**Collocations** :
- *tendance food* (10) \(\Rightarrow\) N *food*(ADJ)
- *phénomène food* (9) \(\Rightarrow\) N *food*(ADJ)
- *projet food* (5) \(\Rightarrow\) N *food*(ADJ)
**Det (masc) food* (10) \(\Rightarrow\) *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%) and 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.
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

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, i.e. 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)

<table>
<thead>
<tr>
<th>Affix</th>
<th>Nb</th>
<th>Exemples</th>
</tr>
</thead>
<tbody>
<tr>
<td>cyber-</td>
<td></td>
<td>cybercondriaque, cyberathlète, cyberattaquer</td>
</tr>
<tr>
<td>e-</td>
<td></td>
<td>e-citoyenneté, e-enseignant, e-recruter</td>
</tr>
<tr>
<td>bio-</td>
<td></td>
<td>bio-exorciste, affinité, bio-diversifier</td>
</tr>
<tr>
<td>éco-</td>
<td></td>
<td>éco-jardin, éco-touristique</td>
</tr>
</tbody>
</table>

- **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-Ving (car-jacking, home-staging, speed-dating, speed-watching, binge-viewing, ride-sharing).
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.
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.