Diachronic Semantic Evolution Automatic Tracking: a pilot study in Modern and Contemporary French combining Dependency Analysis and Contextual Embeddings

Emmanuel Cartier, Université Sorbonne Paris Nord, LIPN UMR7030 CNRS
emmanuel.cartier@univ-paris13.fr
Contents

- Context and Research Questions
- Methodology
- Results & Analysis
- Conclusion / Perspectives
Lexical semantic change (LSC) : omnipresent, multidimensional

- Evidence of lexical change
  - variation of meaning across linguistic communities : location (diatopy : char), sociolinguistic parameters (diastraty : covidiot), pragmatic parameter (communication settings)
  - variation through time : voiture, téléphone...
  - Metaphorical / metonymical use : Mannheim is a (whatever except a town!)
- Formal, phrasal and semantic neologisms
- Stability of meaning remains the base : we manage to communicate!
- Of paramount importance for lexicography : languages are dynamic systems

- Research questions
  - how to model LSC and its parameters? What hints can we use to detect semantic shifts?
  - how to track LSC in the numerical era, with the availability of more and more corpora and language use data?
Context

- Current project (funded by Labex Empirical Foundations of Linguistics, Paris) - EvolSem
  - Setup a reference dataset of semantic lexical change in Modern and Contemporary French (for Linguistic, Computational Linguistics research ... and the general public)
  - Setup a web exploration platform to get CL state-of-the-art approaches to semantic meaning and semantic change detection
  - Setup a web platform to retrieve and edit potentially evolving lexical units
  - Prototype a automatic system to detect and track lexical semantic change
- Here I present the first steps of this project
Previous works

- **Research on lexical semantics and lexical semantic change**
  - Etymology: reconstructing the origin of words and meanings
  - Traditional linguistics: extension/restriction, metaphor, metonymy, denotation/connotation, polysemy (Bréal, 1897)
  - Cognitive linguistics / semantics: metaphor and analogy at the root of human categorization / prototypes and peripheral uses, entrenchment as the process of use emergence, diffusion and adoption (Langacker, 1989; Schmid, 2020)
  - Corpus linguistics / distributional linguistics / distributional semantics (Baroni and Lenci, 2010):
    - First-order principle: « you shall know a word by the company it keeps, » (Firth, 1957)
    - Second-order principle: « if we consider words or morpheme A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution » (Harris, 1954)
  - Sociolinguistics: dia- parameters of variation (eg Coseriu, 1955)
  - Pragmatics: discourse and communication events maintain and modify the evolving system

- **Two tracks of research are studied here, each derived from the distributional hypothesis**
  - Patterns of usage / meaning and their change (collocations, collostructions, behavioral profile (Gries, 2012)
  - Word and Contextual embeddings as an approximation of the distributional hypothesis
Methodology: EvolSem project

- Diachronic corpora
- Choice of evolving lexemes
- Automatic tools: dependency analysis of sentences and exploration of patterns of usage
- Automatic tools: word embeddings and contextual embeddings
- Annotation of sentences / meaning or similarity

Diagram:
- Corpora
  - Automatic retrieval of sentences
  - Dependency analysis
  - Word and contextual embeddings
- Lexemes
  - Exploration tools
  - Annotation tool: meaning / sentence
- Reference dataset!
Methodology: diachronic corpora

- Project Focus
  - Timeframe: modern and contemporary French: 1800 - now
  - Genre: general language > general press

- Diachronic corpora in French: rare and quantitatively small (or noisy!)
  - Google Books NGrams (Michel et al., 2010)
  - Frantext: 1,152 documents (1800-1900), essentially from literature
  - Gallica (Bibliothèque Nationale de France): the most significant and complete resource even if not focusing on corpus linguistics, availability through an API, complex queries, free-of-charge (up to 1950 due to copyright constraints) - low OCR quality for old documents (up to 1850)

- Contemporary corpora: a variety of choices!
  - JSI Timestamped corpora (Trampus et al., 2012): monitor corpora from RSS feeds (2014-), available through SketchEngine API, retrieve sentences and pos-tagging analysis

- Outcome
  - Two periods: Gallica (1800-1850) and JSI (2014-2020)
  - Quantity: at least 200 sentences per period for each evolving lexeme (up to 5000)
Methodology : choice of lexemes

- **(Research) Question**: how to find words that have undergone a semantic evolution, systematically?
  - Absence of (systematic) lexicographic work focusing on LSC => setup from scratch or from etymological dictionaries?

- **Available methods**
  - Lexemes with frequency shifts through time : from Google Ngrams (noisy)
  - Start from the assumption : polysemy as the result of semantic changes (e.g. Bybee, 2015) - requires an available (and freely accessible) database of words
  - X-WIC (Raganato et al., 2021) reference dataset : retrieve from wiktionary all words with several meanings, with the definition and illustrative contexts => automatic approach enables to create a « big » reference dataset (but binary similarity judgements between meanings)

- **Method used in the project**
  - French Wiktionary as a valuable resource (moderation, mostly inspired by reputed dictionaries with a continuous update of meanings / words) - not perfect, but a good base
  - Retrieve set of nouns and verbs having at least two senses and having an “obsolete” mark, for at least one of the senses : 21,837 nouns and 7,997 verbs
  - Filter : at least 200 phrasal contexts in the Gallica press corpus and in the JSI corpus : 13,502 nouns and 5,187 verbs.
  - From these candidate lexemes, sample of 100 verbs and 100 nouns
Methodology: dependency analysis

- Goal: from a sample of sentences
  - extract lexico-syntactic patterns of use
  - explore the link between patterns and meaning

- Procedure
  - UDPipe analysis (Straka, 2018, [https://ufal.mff.cuni.cz/udpipe/2](https://ufal.mff.cuni.cz/udpipe/2))
  - Aggregation of patterns according to syntactic valid patterns (e.g., for Nouns: N + ADJ, N de N, V + N (as object), N (as object) + V)
Methodology: word and contextual embeddings

- **Goal**
  - Check change of meaning with change of similar words

- **Procedures**
  - **Word embeddings**: Word2vec (Mikolov et al., 2013), FastText (Bojanovski et al., 2016) pretrained language models for French: retrieve most similar words of a given word (no polysemy handling). Word2vec model trained with Gallica news press corpora.
  - **Contextual embeddings**: French version of BERT, CamemBERT (Martin et al., 2019), pretrained => retrieve most probable words in sentences with masked language prediction task, and aggregate results. No available model for first period.

**Aggregation**

<table>
<thead>
<tr>
<th>Group</th>
<th>count(lexe...)</th>
</tr>
</thead>
<tbody>
<tr>
<td>voiture</td>
<td>1421</td>
</tr>
<tr>
<td>maison</td>
<td>106</td>
</tr>
<tr>
<td>moto</td>
<td>88</td>
</tr>
<tr>
<td>véhicule</td>
<td>44</td>
</tr>
<tr>
<td>chambre</td>
<td>40</td>
</tr>
<tr>
<td>machine</td>
<td>28</td>
</tr>
<tr>
<td>personne</td>
<td>24</td>
</tr>
<tr>
<td>voitures</td>
<td>23</td>
</tr>
<tr>
<td>ville</td>
<td>21</td>
</tr>
<tr>
<td>femme</td>
<td>20</td>
</tr>
<tr>
<td>route</td>
<td>18</td>
</tr>
<tr>
<td>bus</td>
<td>17</td>
</tr>
</tbody>
</table>

English: for more than twenty years now, he has been defending its commitment to the electric car.
Current Results

- Exploration platform
  - dependency analysis
  - Word embeddings and contextual embeddings
- Case-studies: réaliser / téléphone / glaner
Results: dependency analysis (réaliser)

- **Make sthg become real (achieve sthg):** they made a green building (a green building was realized) / Ils ont réalisé un bâtiment écologique
- **Happen / become true:** if somebody’s fear are realized… / l’impossible s’est réalisé
- **Become aware of sthg:** I realize my error, I realize that… / je réalise mon erreur / je réalise que…

**Common:**
- **Make sthg become real (achieve sthg):** X (agent) réaliser NOUN [any tangible object or that can have a concrete form [dream]] : réaliser des économies, une réforme, un édifice etc.
- **Common:** **Happen / become true:** NOUN se réaliser
- **JSI:** **Become aware of sthg:** X réaliser que / X réaliser NOUN (evaluation/judgement on a fact/situation)
Results: embeddings (réaliser)

- **Make sthg become real (achieve sthg):** they made a green building (a green building was realized) / Ils ont réalisé un bâtiment écologique
- **Happen / become true:** if somebody’s fear are realized... / l’impossible s’est réalisé
- **Become aware of sthg:** I realize my error, I realize that... / je réalise mon erreur I realize that...

---

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Il réalise son rêve de succès aux élections</td>
<td>Il est un édifice écologique de très haute qualité</td>
</tr>
<tr>
<td>Il concrétise son rêve de succès aux élections</td>
<td>Il constitue un édifice écologique de très haute qualité</td>
</tr>
<tr>
<td>Il exprime son rêve de succès aux élections</td>
<td>Il construit un édifice écologique de très haute qualité</td>
</tr>
<tr>
<td>Il poursuit son rêve de succès aux élections</td>
<td>Il possède un édifice écologique de très haute qualité</td>
</tr>
<tr>
<td>Il affiche son rêve de succès aux élections</td>
<td>Il offre un édifice écologique de très haute qualité</td>
</tr>
</tbody>
</table>

---

- **He realizes his dream of success in the elections:** il réalise son rêve de succès aux élections
- **It creates a high quality ecological building:** il est un édifice écologique de très haute qualité

---

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Il dit que son projet n'est pas réaliste</td>
<td>Il précise que son projet n'est pas réaliste</td>
</tr>
<tr>
<td>Il estime que son projet n'est pas réaliste</td>
<td>Il affirme que son projet n'est pas réaliste</td>
</tr>
<tr>
<td>Il déclare que son projet n'est pas réaliste</td>
<td>He realizes that his project is not realistic</td>
</tr>
</tbody>
</table>
Results: dependency analysis (téléphone)

- **Verbal patterns (Noun (subject) Verb)**
  - common features: sonner, transmettre,
  - new features: denoting some semantic features of the new meaning (vibrer, biper, borner, etc.)

- **Modifier patterns (Noun ADJ)**
  - new features: téléphone mobile (> mobile), téléphone portable (> portable), téléphone intelligent versus smartphone (> smartphone)
  - Current concurrency - téléphone versus portable versus smartphone (Metropolitan France) versus cell phone (Canada) / handy

Dependency analysis as a way to access semantic features
Results : embeddings + dependency analysis (gplaner)

1800-1850 (by frequency of usage)
1/ semer/récolter [to sow/harvest] (gplaner des épis de blé [to glean ears of corn]),
2/ lire/apprendre [to read/to learn] (gplaner après les anciens [to glean from the elderly]),
3/ trouver/chercher/découvrir [to find/search/discover] (gplaner des informations [to glean information])

2014-2020
3/ trouver/chercher/découvrir [to find/search/discover] (gplaner des informations [to glean information])
4/ gplaner/gagner [to glean/to win] (gplaner des trophées [to win trophies])
2/ has disappeared

=> fill-mask task allows to discover, by induction, the semantic features that explain the meaning shifts.
The historically primary meaning of gplaner denotes the **collection, after the main harvest, of the remains of the ears of wheat or any other crop.**

**Two semantic features thus coexist:**
- the **remainder**, and at the same time a **remainder of a certain value** (food at first).
The conjunction of these two traits is preserved by metaphor in senses 2 and 3, while only the second trait is used in sense 4, essentially applied to sports.
Conclusion / perspectives

- Pilot study to identify semantic structure and its evolution from three parameters
  - (Cognitive aspects) frequency change,
  - (Linguistic aspects) patterns of usage, distribution similarity,
- About 10 nouns and 10 verbs with meanings and prototypical sentences
- Exploration web platform will be available soon
Conclusion / perspectives

- **Contextual Embeddings**
  - **Advantages**
    - Useful to identify meanings by providing synonyms
    - most of the time, all meanings are present but with a different importance in use (eg: glaner => récupérer dans les champs versus recueillir une information versus gagner qch en sport)
  - **Drawbacks**
    - Very sensitive to sentence structure => requires simple sentences and even prototypical sentences as meaning anchors
    - Mix of synonyms, hypernyms and co-hyponyms
Conclusion / perspectives

- Patterns of usage / dependency analysis
  - Advantages
    - Useful to get prototypical patterns of usage and the paths of lexical innovation - readable and interpretable
    - Combination of syntactic patterns ans lexico-syntactic patterns the most promising avenue of research
    - New meanings appear first explicitly (example : téléphone mobile, portable) before being embedded in a short form (>téléphone).
    - Patterns can also retrieve (semi-)frozen multiword expressions
  - Drawbacks
    - Very sensitive to sentence structure => requires simple sentences
    - Association measures to the rescue!
Conclusion / perspectives

- **Diachronic Lexical Semantic Change Detection still in its infancy**
  - State-of-the-art systems only work on toy reference datasets, mainly in English
  - Current monopoly of Neuronal Approaches, whereas these methods are approximated numerical representation of linguistic features

- **Need for a combination of criteria:**
  - Frequency shift (as the main hint of entrenchment)
  - Usage patterns shift
  - Distributional shifts
  - Sociolinguistical / contextual shifts

- **Next steps:**
  - Manual choice of prototypical sentences as anchor for meaning ans clustering of Embeddings to show evolution and graduality
  - Usage patterns complemented with association measures to get MWE and the most accurate new patterns
  - Annotation campaign of sentences similarity to check the expert annotations

- Do not overtrust corpora but use them!
- Thank you!
Methodology: linguistic annotation

- Two schemes: discrete meanings versus prototypical meanings and peripheral uses

<table>
<thead>
<tr>
<th>Method / references</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>a priori sets of meaning - annotators decide on meaning annotation per sentence + inter-annotator agreement</td>
<td>Explicit meanings</td>
<td>Arbitrary set of meanings Low inter-annotator agreement</td>
</tr>
<tr>
<td>retrieve sample of sentences with the given word - annotators are presented pairs of sentences and they decide on the similarity of meanings + inter annotator agreements</td>
<td>Higher inter annotator agreement Gradability of meanings and prototypical meanings</td>
<td>Implicit meanings</td>
</tr>
</tbody>
</table>

Advantages
- Explicit meanings
- Higher inter annotator agreement
- Gradability of meanings and prototypical meanings

Drawbacks
- Arbitrary set of meanings
- Low inter-annotator agreement
- Implicit meanings
Methodology: linguistic annotation

- **Linguistic description of meanings with several instructions**
  - Prototypical versus peripheral meanings
  - Describe prototypical pattern uses for every meaning
  - Describe lexical change processes
  - Find prototypical sentences