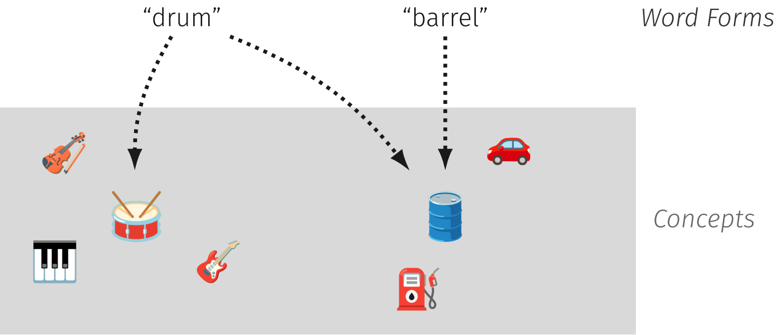
Towards an Onomasiological Study of Lexical Semantic Change through the Induction of Concepts

Bastien Liétard, Mikaela Keller, Pascal Denis

August 15, 2024

Context

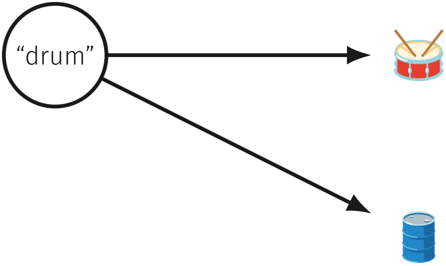
The word-meaning mapping



The Semasiological View of Lexical Semantic Change

Meaning of "drum" before the 19th century.

The Semasiological View of Lexical Semantic Change



Meaning of “drum” **after** the 19th century.

NLP approach for Semasiological LSC

1. Compute a **representation** for each occurrence $o^T_{w,i}$ of word w at time T.
   $\rightarrow$ typically using a Contextualised Language Model.

NLP approach for Semasiological LSC

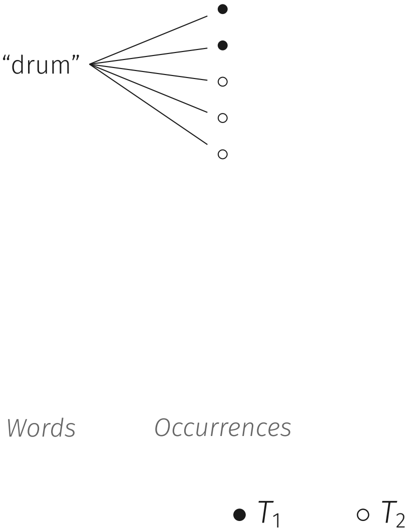
1. Compute a **representation** for each occurrence $o_{w,i}^{T}$ of word w at time T.
   $\rightarrow$ typically using a Contextualised Language Model.
2. For each time period T, cluster occurrences of w into **sense clusters** to assign to each occurrence a *sense*.

NLP approach for Semasiological LSC

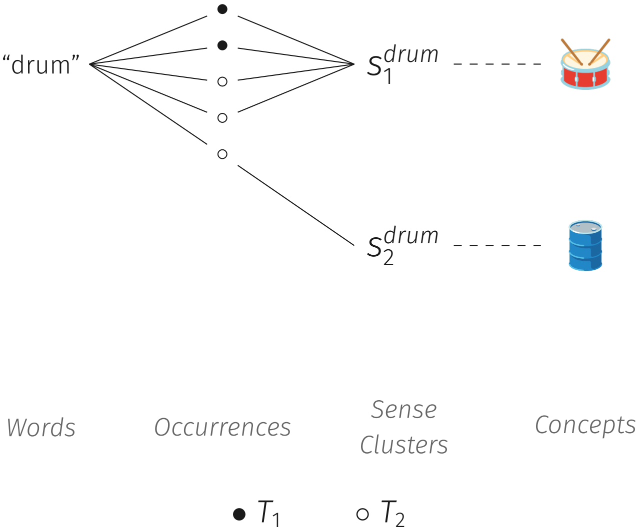
1. Compute a **representation** for each occurrence $o^T_{w,i}$ of word w at time T.
   $\rightarrow$ typically using a Contextualised Language Model.
2. For each time period T, cluster occurrences of w into **sense clusters** to assign to each occurrence a *sense*.
3. **Compare the sense distributions** between pairs of periods T_1 and T_2 (assuming we can *align clusters* from different periods).
   $\rightarrow$ typically using a metric like the *Jensen-Shannon Divergence*.

(Giulianelli et al., 2020; Martinc et al., 2020)

NLP approach for Semasiological LSC (ctd.)

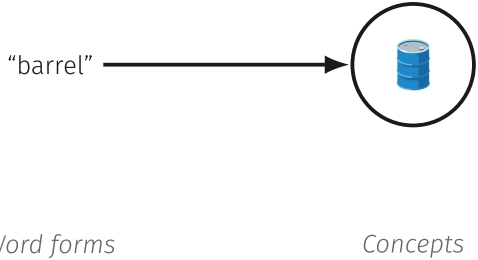


NLP approach for Semasiological LSC (ctd.)



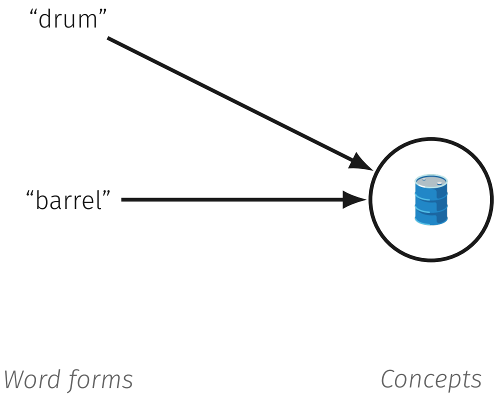
Inducing Concepts for Onomasiological LSC

The Onomasiological View of Semantic Change



Naming of the concept of *large cylindrical container for liquids* before the 19th century.

The Onomasiological View of Semantic Change



Naming of the concept of *large cylindrical container for liquids* **after** the 19th century.

Inducing Concepts with Double-Clustering

1. Compute a **representation** for each occurrence $o^T_{w,i}$.
2. For each time period T, cluster occurrences of w into **sense clusters** to assign to each occurrence a *sense*.

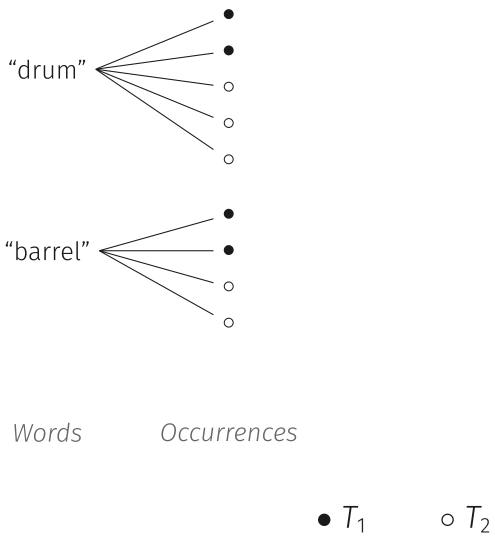
Inducing Concepts with Double-Clustering

1. Compute a **representation** for each occurrence $o^T_{w,i}$.
2. For each time period T, cluster occurrences of w into **sense clusters** to assign to each occurrence a *sense*.
3. Cluster the sense clusters into **concept clusters**, to assign each occurrence to a *concept*.

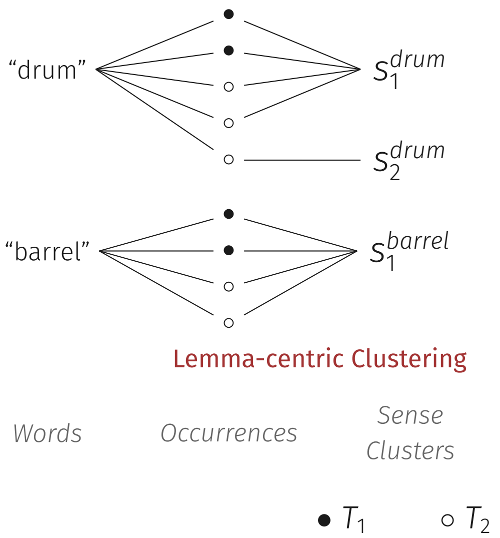
Inducing Concepts with Double-Clustering

1. Compute a **representation** for each occurrence $o_{w,i}^{T}$.
2. For each time period T, cluster occurrences of w into **sense clusters** to assign to each occurrence a *sense*.
3. Cluster the sense clusters into **concept clusters**, to assign each occurrence to a *concept*.
4. Compare **how concept distributions** (of words) or **word inventories** (of concepts) changed, assuming we can align concept clusters over time.

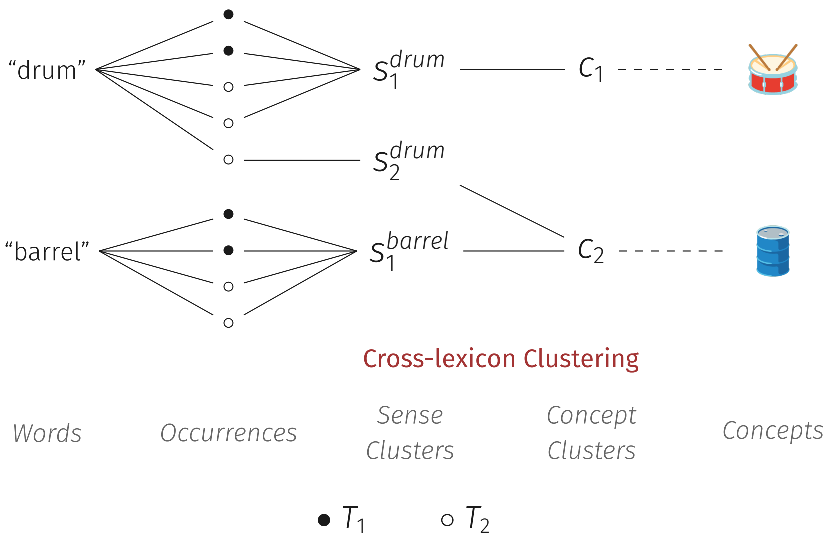
Inducing Concepts with Double-Clustering (ctd.)



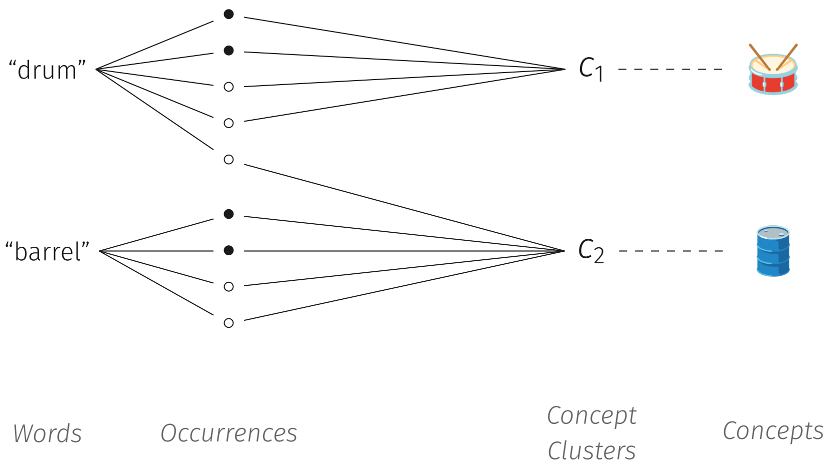
Inducing Concepts with Double-Clustering (ctd.)



Inducing Concepts with Double-Clustering (ctd.)



Inducing Concepts with Double-Clustering (ctd.)



● T_1 ○ T_2

Key Components

1. A **target lexicon** and the corresponding occurrences.
2. A **representation mode** for occurrences.
3. A **lemma-centric clustering** algorithm applied to *occurrences*.
4. A **cross-lexicon clustering** algorithm applied to *sense clusters*.
5. A **temporal cluster alignment** strategy.

Experimental Application

Setup

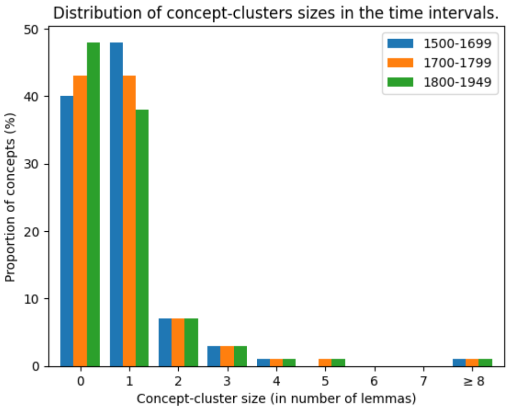
- **Data:** PRESTO (core) Corpus (French, 1500-1950).
  53 texts, various genres, fiction and non-fiction.
- **Target lexicon:** 623 nouns.
  314K occurrences total, occs/word ratio : 504.4
- **Time periods:** 1500-1699, 1700-1799, 1800-1949 (balanced in # of occurrences).
- **Representation mode:**
  - Partial sentence lemmatization (N,V,ADJ, ADV)
  - Embeddings from hidden-layers of XLM-R (Large).

Clustering Strategy

- **Lemma-centric clustering:**
  - Hierarchical Agglomerative (linkage: minimum).
- **Cross-lexicon clustering:**
  - Averaging occurrence vectors in sense clusters
  - Hierarchical Agglomerative (linkage: average).
- **Cluster alignment:** All periods are mixed together during clustering.
- **Algorithms / Hyperparameters selection:** Highest amount of concept clusters containing 2-5 distinct words.

Statistics on Obtained Concept Clusters

867 concept-clusters across the 3 time periods; Only 265 (31%) appear in all 3 periods.



The Single-Word Clusters

In each period, 40% of concept clusters **contain only 1 unique target word**.
It includes 51% of the 265 concept clusters instantiated in all 3 periods.

Clark (1993)'s **Principle of Conventionality**:

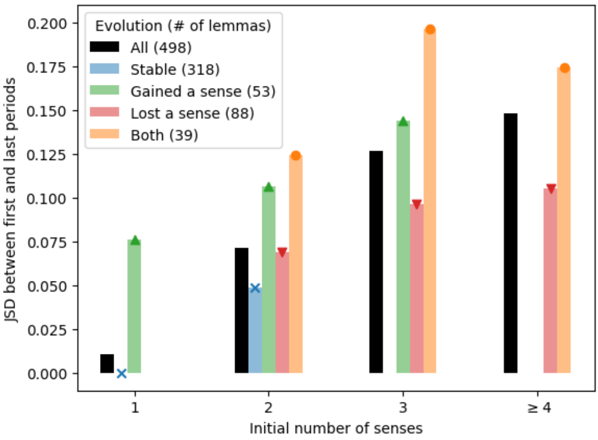
For certain meanings, there is a form that speakers expect to be used in the language community.

Quality of Obtained Clusters in 1800-1949

		Cluster size		
Category	Total	2	3	4
Nb. of clusters	101	62	29	10
Synonyms	27%	32%	24%	0%
Near-synonyms	20%	15%	28%	30%
Lexical / topical relations	40%	42%	38%	40%
Invalid cluster	13%	11%	10%	30%

Table 1: Categorization of small induced concept-clusters in 1800-1949. Invalid clusters are those showing no semantic relation.

Semasiological Lexical Semantic Change



Significant correlation:
- initial # of senses
- JSD

(Law of Innovation, Hamilton et al. 2016; Luo and Xu 2018)

Onomasiological Lexical Semantic Change

Concept Evolution	#Concepts
Expanded naming	27 (10%)
Shrinked naming	5 (2%)
Both	6 (2%)
Identical naming	227 (86%)

Expanded naming: {"peuple", "tribu"} (PEOPLE), {"feu", "incendie"} (BIG FIRE).
Shrinked naming: {"pourquoi", "parquoi"} (EXPLANATION), {"amour", "amitié"} (ROMANTIC LOVE).

Conclusion and Perspectives

In this talk, we introduced a methodology...

- **inducing concepts** from word occurrences,
- with **no requirement** of predefined concepts,
- allowing both **semasiological** and **onomasiological** studies of LSC.

Perspectives:

- More advanced bi-level clustering strategy
- Different time granularity
- Cluster interpretation

References

E. V. Clark. *Conventionality and contrast*, page 67–83. Cambridge Studies in Linguistics. Cambridge University Press, 1993. doi: 10.1017/CBO9780511554377.005.

M. Giulianelli, M. Del Tredici, and R. Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL `https://aclanthology.org/2020.acl-main.365`.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL `https://aclanthology.org/P16-1141`.

Y. Luo and Y. Xu. Stability in the temporal dynamics of word meanings. In *CogSci*, 2018.

M. Martinc, S. Montariol, E. Zosa, and L. Pivovarova. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3382186. URL https://doi.org/10.1145/3366424.3382186.