A METHOD TO AUTOMATICALLY IDENTIFY DIACHRONIC VARIATION IN COLLOCATIONS

Marcos Garcia, Marcos García-Salido
(LyS Group, CITIC, Universidade da Coruña)
marcos.garcia.gonzalez@udc.gal

1. COLLOCATIONS

- Collocations are syntactically related pairs of lexical units.
  - BASE: freely selected (due to its meaning).
  - COLLOCATE: selection restricted by the base.
- The meaning of the collocate depends on the base:
  - *take*$_{\text{COLLOCATE}}$ *[a] walk*$_{\text{BASE}}$
  - *meet*$_{\text{COLLOCATE}}$ *[a] requirement*$_{\text{BASE}}$
  - *fresh*$_{\text{COLLOCATE}}$ *water*$_{\text{BASE}}$

2. VARIATION IN COLLOCATIONS

- Findings in historical linguistics (Spanish):
  - $hacer_C$ $vergüenza_B \rightarrow dar_C$ $vergüenza_B$
    "[to] cause shame"
  - $poner_C$ $miedo_B \rightarrow dar_C$ $miedo_B$
    "[to] cause fear"
- Understudied area in linguistics and NLP.
- Important to understand language change and to analyze historical corpora.

3. GOALS

➔ Assist historical linguists.

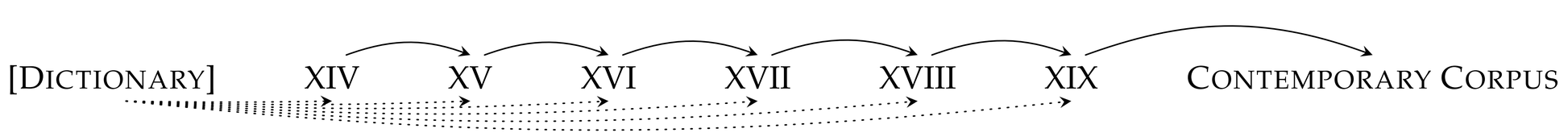
➔ Track collocations over time.

➔ Identify types of variation
(4 types + 3 frequency trends).

➔ Search for alternatives with similar meaning.

4. METHOD

1. Select candidate collocations:
   dependency parsing + statistical measures.
2. Analyze each candidate in the following periods.
3. Classification: loss of the collocation (1), base (2), collocate (3), or combination (4) [+ 3 frequency trends].
4. Search for alternatives (new collocate, verbalization).



- Resources: diachronic corpora + hist. embeddings.
- Optional: contemporary corpus + dictionary.

5. EXPERIMENTS AND RESULTS

- In Portuguese and Spanish.
- Normalized and *noisy* corpora.
- Verb-object collocations.
- LinguaKit + UDPipe + *fasttext*.
- Quantitative: 69% precision.
- Qualitative (some examples):

  Pt: *deitar→dizer missa* (XVI).
  "[to] say [a] mass"

  Pt: *dar [um] alegrão → alegrar* (XVII).
  "[to] make happy"

  Es: *(a)prestar→tener paciencia* (XVIII).
  "[to] have patience"

  Es: *meter→poner paz* (XVII).
  "[to] put peace"

6. CONCLUSIONS

- Simple but useful method.
- Identify 4 change types.
- Further work:
  - Adapted NLP tools.
  - Contextual. embeddings.
  - Visualization tool.