LINKING VERB PATTERN DICTIONARIES OF ENGLISH AND SPANISH

VÍT BAISA
Masaryk University
Brno, Czech Republic

~

SARA MOŽE
University of Wolverhampton
United Kingdom

~

IRENE RENAU
Pontificia Universidad Católica de Valparaíso, Chile
INTRODUCTION

• Verbs are complex
• **AIM**: methodology and tools for the creation of a multilingual corpus-driven lexical resource for verbs using manual and automatic procedures
• CPA-based monolingual pattern dictionaries
  – What are they?
• New multilingual resource – researchers and language professionals?
• Preliminary study:
  I. **Manual linking task** \(\rightarrow\) gold standard dataset
  II. **Automatic linking task** = algorithm; evaluated against the gold standard
CORPUS PATTERN ANALYSIS (CPA)

• Corpus Pattern Analysis (CPA) (Hanks, 2004)
  – an empirical technique in Corpus Ling. and Lexicography
  – map word meaning onto word use through lexical analysis of phraseological patterns, collocations

• Basis: Theory of Norms and Exploitations (TNE) (Hanks, 2013)
  – ‘double helix‘ – patterns of normal usage (‘norms’) vs. their ‘exploitations’

• ‘Pattern‘ – semantically motivated syntagmatic pattern
  – Syntax: SPOCA (Halliday)
  – Semantics: typical nominal slot fillers, represented by Semantic Types (ST) – mnemonic sem. labels
    ▪ CPA shallow ontology (Hanks and Ježek, 2010) – approx. 250 STs; shared by several projects
WHAT IS A PATTERN?

• PDEV: harvest

<table>
<thead>
<tr>
<th>#</th>
<th>%</th>
<th>Pattern &amp; primary impliciture</th>
</tr>
</thead>
</table>
| 1  | 81.11%| *[[Human]]* harvest *[[Plant = Crop]]*  
*([Human]) cuts down and gatherings *[[Plant = Crop]]*  
when *[[Plant]]* is ready for use |
| 2  | 5.00% | *[[Human]]* harvest *[[Location]]*  
*([Human]) gathers foodstuff from *[[Location]]* |
| 3  | 11.11%| EUPHEMISM *[[Human]]* harvest *[[Fish | Animal]]*  
*([Human]) kills *[[Fish | Animal]]* for use as food |
| 4  | 2.78% | BIOCHEMISTRY, JARGON *[[Human]]* harvest *[[Body_Part]]*  
*([Human]) removes *[[Body_Part]]* for research or transplanting |
CPA PATTERN DICTIONARIES

• Pattern Dictionary of Italian Verbs (PDIV) – Elisabetta Ježek, Pavia

• **Pattern Dictionary of English Verbs (PDEV)**
  
  
  – Prof. Hanks, University of Wolverhampton; over 1,700+ English verbs completed
  
  – Procedure: corpus samples (250/500/1000 lines) from the BNC corpus (Leech, 1992);
    
    • Sketch Engine – word sketches (Kilgarrif et al., 2014),
    
    • CPA Editor (Baisa et al., 2015) and CPA shallow ontology (Ježek and Hanks, 2010)
    
    • Implicatures; register, domain, idiom/phrasal verb labels; links to FrameNet (Ruppenhofer et al., 2010)
    
    • Percentages for each pattern

• Pattern Dictionary of Spanish Verbs (PDSV)
  
  
  – Verbario: Irene Renau, Pontificia Universidad Católica de Valparaíso
  
  – 300 high-frequency Spanish verbs (currently only 100 publicly available online)
  
  – Same methodology (CPA), guidelines, ontology, tools (SkE); but: Spanish Web Corpus
MANUAL LINKING: SP-EN PATTERN PAIRS

• Gold standard:
  – 87 SP verbs with one or more EN equivalents (total: 126 EN verbs)
  – Medium-frequency verbs, up to 15 patterns
  – Manual cross-linguistic links between pattern pairs
    ▪ **semanto-syntactic similarity** = tertium comparationis
    ▪ linking procedure developed
    ▪ dataset used in algorithm evaluation

• Issues – practical, theoretical
  – **Coverage**: PDEV/PDSV are WIP resources; different coverage;
    → **limited overlap!!!**
  – Zero equivalence: cultural, social, cognitive, pragmatic reasons; idioms
INPUT: POTENTIALLY MATCHING EN PATTERN

Does it have the same basic syntactic structure as the SP pattern (i.e. SVO or SV [+no obj])?

**YES**

Do all semantic types in all obligatory syntactic slots match? E.g.: EN: [[Human]] admire [[Anything]]
SP: [[Human]] admirar [[Anything]]

**YES**

**OUTPUT: PERFECT MATCH**

**NO**

**OUTPUT: NO MATCH**

**NO**

Do the two patterns share at least ONE semantic type in the same obligatory syntactic slot? For example:
EN: [[Eventuality 1 | Human | Institution]] occasion [[Eventuality 2]]
SP: [[Eventuality 1]] motivar [[Eventuality 2]]

**YES**

**OUTPUT: PARTIAL MATCH**

**NO**

Are the two semantic types in the same obligatory syntactic position related to each other in terms of inheritance in the CPA ontology (up to two nodes), e.g. [[Eventuality]] (supertype) vs. [[Activity]] and [[Plan]] (subtypes):
EN: [[Eventuality 1 | Human]] spoil [[Eventuality 2]]
SP: [[Eventuality | Human]] estropear [[Activity | Plan]]

**YES**

**OUTPUT: PARTIAL MATCH**

**NO**

**OUTPUT: NO MATCH**
AUTOMATIC PATTERN LINKING: ALGORITHM

• **Heuristic-based algorithm:** automatic linking suggestions

• **Similarity score**
  – 490 SP patterns and their translations into EN (statistical EN-SP dictionary <-- parallel corpus)
  – S, DO, IO → comparison of STs
  – **Full match:** 1 score pt (*Human = 0.5 pt); matching empty slots (e.g. DO) – 0.5 pts
  – **CPA ontology:** similarity score = 0.5\(^N\)
    ▪ Score calculated based on the distance (N) in the CPA ontology tree
  – Scores summed up, final score assigned to the pair, top ranking EN pattern = most likely candidate

• **Evaluation**
  – 50 SP-EN *verb* pairs
  – Excluded: SP pattern cannot be matched against an EN pattern in the sample
  – Final no. of candidate pattern pairs: 50 → gold standard
  – 40/50 suggested candidate pairs were correct → **80% precision**
CONCLUSION

• Future activities:
  – **Gold standard**: more annotated data;
  – Refine the linking procedure (fine-grained distinctions?; intralingual links)
  – **Algorithm**: train, improve precision;
  – **Software adaptation**: feature for adding cross-linguistic links to the dictionaries/databases.
REFERENCES

USEFUL LINKS

• Pattern Dictionary of English Verbs
  http://pdev.org.uk/

• VERBARIO (Pattern Dictionary of Spanish Verbs)
  http://www.verbario.com/

• PDEV-LEMON
  http://pdev.org.uk/PDEVLEMON.html