**Object of study**

- We focus on ways of automatically retrieving compound dictionaries from sentence-aligned corpora using WordNet for the pair of languages German—Spanish.
- German—Spanish compound correspondences are of the type 1:n:
  
  (1) Warm Wasser → ‘Preparación de agua caliente’
  (2) Wärme Rügewinnung → ‘sistemas de recuperación de calor’

- The ultimate aim is to integrate the extracted compound dictionaries in Statistical Machine Translation (SMT) tasks.

**Gold Standard**

Our Gold Standard consists of 168 compounds and their translations:

- They were extracted from the TRIS corpus [1], a specialised German—Spanish corpus.
- All compounds were split and tagged with their corresponding Part-of-Speech (PoS) tags [2].
- All translation correspondences were also PoS tagged [2].
- If a compound had several translation correspondences, each was stored as a different entry in the Gold Standard.

**Compound-Phrase Matching**

1. Given a split German compound C, there is a list of lemmas \(C = \{c_0, ..., c_n\}\).
2. Given a Spanish sentence aligned to the German sentence that contains C, there is a list of lemmas \(S = \{s_0, ..., s_n\}\).
3. Be type(x) a function that retrieves the semantic type of a word, obtained from Wordnet.
4. For each German compound, Spanish sentence pair \((C,S)\):
   (a) Locate the translated root of C in S by finding a lemma \(s_x\) in S with a semantic type that matches the root of the compound, i.e. \(\text{type}(s_x) = \text{type}(c_0)\).
   (b) Locate the rightmost word in the Spanish phrase that translates C by finding a lemma \(s_y\) in S with a semantic type that matches the first lemma of the compound, i.e. \(\text{type}(s_y) = \text{type}(c_0)\).
   (c) The candidate Spanish phrase that translates C is the span of words defined as \([s_x, ..., s_y]\).

**Challenges faced**

- PoS taggers: More damaging on the Spanish side when not locating phrase roots.
- WordNet coverage.
- Manual semantic matching: GermaNet has a potentially useful adjective classification that maps unevenly to the Spanish WordNet.

**Working hypothesis: Semantic Types Mapping**

Our working hypothesis is that different formants of a compositional compound will share semantic features with their corresponding translational equivalents:

**Conclusion and Future Work**

- Expand the Gold Standard.
- Evaluate the PoS tagger and identify sources of error that might be avoided. Eventually test other PoS taggers.
- Redefine the type(x) function to make it not only dependent on the first listed sense of each WordNet.
- Align semantic classes automatically using word-alignment techniques, or using the English WordNet as a pivot.
- Use supervised machine learning to predict Spanish phrase spans from the German compounds.

**References**