# Extraction of Multilingual MWEs from Aligned Corpora

Eric Wehrli[♢] Aline Villavicencio[♣]

[♢] University of Geneva (Switzerland) and [♣] Federal University of Rio Grande do Sul (Brazil)

Eric.Wehrli@unige.ch, avillavicencio@inf.ufrgs.br

## Summary and Goals

On-going research for establishing a sizeable **multilingual repository of MWEs extracted from aligned corpora**, for **Portuguese, English and French**. The MWE types include **collocations, named entities, compounds, idioms**.
In a first phase, MWEs will be extracted monolingually from the aligned sentences in Europarl corpus, and in a subsequent phase to other bilingual or multilingual corpora, such as the Brazilian newspaper Folha online.

## Method

For MWE extraction we use:

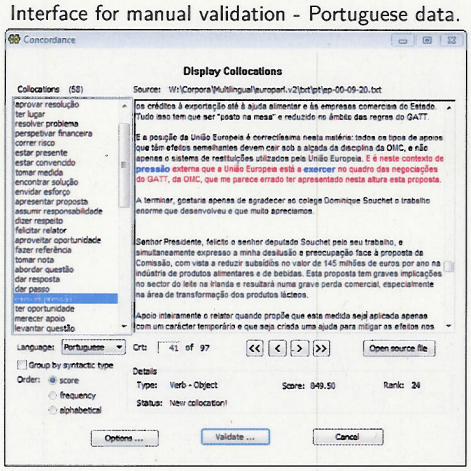
- symbolic parsing-based methods, for extracting collocations and other MWEs corresponding to grammatically well-formed phrases, for precision oriented results;
- statistical methods, for extracting MWEs based on surface proximity, such as compounds, named entities and idioms, using POS patterns and association measures, for recall oriented results.

Poster focuses on the monolingual extractions, where two methods are applied to each language independently. In a multilingual phase we will adopt distributional methods to find MWE translations in the aligned sentences.

## Parsing-based method

1. candidate extraction: the parser is used to extract collocation candidates from pre-defined structural configurations
   - adjective-noun, noun-prep-noun, verb-direct object, etc.
2. statistical validation: the candidates are then evaluated, using log likelihood association measure [Seretan, 2011].
3. manual validation: the results are displayed in a specific user interface, along with the context in which they occur for manual validation for inclusion in the database:

Interface for manual validation - Portuguese data.




## The parser

The Fips "deep" linguistic parser [Wehrli, 2007;Wehrli et al., 2014] computes constituent structures and grammatical functions, handling long-distance dependencies (e,g, in wh-questions and relative clauses). It is available for each of the three languages (French, Portuguese and English).
For instance, the parser finds the Portuguese collocation *tomar uma decisão*, whose equivalent in French is *prendre une décision* and in English *to make a decision*.

1. As decisões são sempre tomadas pela própria Comissão.
   - Les décisions sont toujours prises par la Commission elle-même.
   - Decisions are always made by the Commission itself.

The parsing-based method

- gives very precise results for collocations and other grammatically well-formed expressions,
- but it depends on parsing coverage for a given language
- cannot easily retrieve MWE types which are not grammatically well-formed, such as *by and large*, or French *donnant donnant*
- doesn't work well with named entities

## Statistical Methods

To complement recall we extract MWEs using statistical methods with the *mwetoolkit* [Ramisch, 2015]:

1. candidate extraction: based on the definition of POS tag patterns
2. statistical validation: using a variety of association measures (pointwise mutual information, log lihelihood, $\chi^2$, etc) to identify candidates whose components co-occur more often than chance
3. manual validation.

The top 20 English N+N candidates sorted by association measures

| |
|---|
| Mr. President, Member States, Madam President, United States, Member State |
| climate change, Middle East, labour market, United Kingdom, candidate country |
| Mr. President-in-Office, EU Member, Commission proposal, action plan, Amendment Nos |
| fellow Member, Security Council, Lisbon Strategy, security policy, death penalty |

The top 20 Portuguese N+N candidates sorted by association measures (translations)

| |
|---|
| Mr. Deputy, Madam Deputy, Mr. Commissioner, candidate country, euro zone |
| Mr. President, Madam Commissioner, automobile industry, member state, member country |
| Madam Speaker, automobile sector, candidate state, Mr. Minister, Madam President |
| partner country, automobile vehicle, candidate country, beneficiary country, Mr. Speaker |

## Conclusions and Future Work

In a first stage this extraction is done monolingually, independently for each language:

- an initial set of candidates is obtained for each language, and is automatically validated for each method

In a second stage we will apply alignment and distributional methods to identify bilingual correspondences for the monolingual candidates [Caseli et al., 2010;Laranjeira et al. 2014]. Additionally, we plan to adopt a weighted voting scheme to combine results from parsing and statistical-based methods for the cases where they disagree based on how confident they are about each candidate.

*We thank the support of PARSEME and project FAPERGS-CNRS-INRIA AiM-WEst.*

LATL
LABORATOIRE D'ANALYSE ET DE TECHNOLOGIE DU LANGAGE

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

PARSEME