Research Material for BENELEX:

Actors, terminology, terminological evolution and contrastive terminological context comparison

Wim Peters (wilhelmuswim@gmail.com)
Research Fellow, BENELEX

February 2019

This document accompanies the working paper:
L Parks, W Peters & M Lennan, Guidelines and Codes on the participation of indigenous peoples and local communities of the Convention on Biological Diversity: A comparative analysis using natural language processing.

It documents the research material created in the BENELEX pilot.

As described in our working paper, we focused in the BENELEX pilot on the discourse regarding indigenous peoples and local communities (IPLCs) in the following three documents:

- Akwe:Kon
- Tkarihwaié:ri
- Mo'otz Kuxtal.

This document describes the research material produced with the assistance of legal informatics and used in the scholarly analysis of these documents.
It also describes additional available material that covers COPS 1-13 and can be further explored in future research (see section 6 below).

1) Named Entities: IPLCs and other actors

The following actor types have been identified in the COP1-13 documents. The lists for each actor type are in directory “gazetteer lists”:

- IPLCs
- Academia
- Agriculture
- Bodies related to IPLCs
- CBD
- Citizen bodies
- Civil society/Citizen bodies
- Government institutions
- Industry
- Multilateral state bodies
- NGOs
- Private institutions

2) Terminology

Terms are important concepts for the content of the domain knowledge they cover.
In BENELEX, a set of 909 valid terms was obtained for the COP domain after manual evaluation of automatically extracted COP term candidates.

The spreadsheet "**selected-terms-all-COPs.xlsx**" lists information regarding these approved terms. This list has been used in the analysis of term context in Akwe:Kon, Tkarihwaié:ri and Mo'otz Kuxtal described in the working paper.

a) Tab “selected-terms-all-COPS” lists all terms and their frequency in the COP 1-13 documents.

b) Tab “term-relatedness”, lists semantic relations between terms.
The following relations are used:

- hasSynonym: terms that share the same meaning.
- hasNearSynonym: terms that are very similar but not exactly the same.
- hasBroader: broader terms in a thesaurus have a more general meaning.
- hasNarrower: narrower terms in a thesaurus have a more specific meaning.
- hasRelated: terms that are in some way related through some unspecified relation

c) Tabs 3-15 lists terms found in individual COPS (1-13).

3) Term Evolution

The spreadsheet "**IPLC-context-term-evolution-akwe-tkari-mootz.xlsx**" lists the evolution of terminology seen as timeline ordered transitions of terminology occurring from Akwe:Kon (akwe), through Tkarihwaié:ri (tka) into Mo'otz Kuxtal (mootz).
It provides the following information:
- unique terms per document with frequency
- terms lost/kept/new/regained in each transition with frequency. "regained" is only applicable to Mo'otz-Kuxtal when it re-adopts terms from Akwe:Kon that were lost in Tkarihwaié:ri.
- graphics illustrating the evolution.

4) Term Context Comparison

The spreadsheet "**term-context-for-IPLCs-akwe-kon-mootz-kuxtal-tkarihwaie.xlsx**" offers a pairwise comparison of the three documents, listing:
- the terms occurring in the paragraphs containing mentions of IPLCs with their frequency
- term overlap between documents for IPLC mentions
- context terms unique to each document with their frequency

5) Extension of the contrastive comparison to COP documents 1-13

The spreadsheet "**all-cops-contrastive-terminology-IPLC-Industry.xlsx**" contains the terminology associated with IPLCs and Industry across the whole COP1-13 corpus.
It has a number of tabs in which the automatically extracted context information is presented in two tiers.

The first 5 sheets cover the high level classes IPLC and Industry, and give an overall view of their term context with the accumulated frequency of their co-occurrence within sentences. Sheets 6-8 drill further down into the instances of IPLC vs Industry (i.e. the actual strings contained in the gazetteer lists described in 1) above). This provides more fine-grained information about the contexts of each individual actor in the IPLC/Industry lists. Sheets 6-8 give you therefore a more detailed comparison, but at the cost of data size (they are larger sheets to work through).

Sheets 9-12 capture collocational information regarding IPLC/Industry entities and their terminological contexts. Collocations are statistically derived pairs of IPLC/Industry elements and terms that have a relatedness score based on their statistical distribution in the COP 1-13 texts (962000 word tokens).
The applied co-occurrence score is Pointwise Mutual Information (PMI; see https://en.wikipedia.org/wiki/Pointwise_mutual_information), which has been computed over the whole corpus of 13 COP documents. The score ranges between 0 and 100.
The co-occurrence window is a sentence, which means that only terminology occurring within the same sentence as the IPLC or Industry instance is taken into account.

The last four sheets are described below.

- "PMI-based term overlap": the list of terms that are shared in the contexts of both IPLC and Industry.

- "PMI overlap per actor": the same overlapping terms in column A but now with added information about which IPLC/Industry entity they co-occur with, what the co-occurrence score is (the higher the more significant), and the frequency of the term-actor collocational pair.

IPLC/Industry actors collocating with the term in column A are grouped together in column B and C, separated by "\".
For instance, the term 'adverse impact' collocates with IPLC actor “indigenous peoples and local communities”, with the additional information “#34.34336869740204#5”. The “#” separates actor instances from PMI score (in this example 34.34336869740204) and frequency (5).

- "PMI IPLC only": terms unique to the IPLC context specified per IPLC instance, ordered by PMI score

- "PMI Industry only": terms unique to the Industry context specified per Industry instance, ordered by PMI score