Knowledge Extraction and Representation: the EcoLexicon Methodology

Pilar León Araúz, Arianne Reimerink
Department of Translation and Interpreting, University of Granada
Buensuceso 11, 18002, Granada, Spain
E-mail: pleon@ugr.es, arianne@ugr.es

Abstract

EcoLexicon, a multilingual terminological knowledge base (TKB) on the environment, provides an internally coherent information system which aims at covering a wide range of specialized linguistic and conceptual needs. Knowledge is extracted through corpus analysis. Then it is represented and contextualized in several dynamic and interrelated information modules. This methodology solves two challenges derived from multidimensionality: 1) it offers a qualitative criterion to represent specialized concepts according to recent research on situated cognition (Barsalou, 2009), and 2) it is a quantitative and efficient solution to the problem of information overload.

Keywords: knowledge extraction, knowledge representation, EcoLexicon, multidimensionality, context

1. Introduction

EcoLexicon[1] is a multilingual knowledge base on the environment. So far it has 3,283 concepts and 14,695 terms in Spanish, English and German. Currently, two more languages are being added: Modern Greek and Russian. It is aimed at users such as translators, technical writers, environmental experts, etc., which can access it through a friendly visual interface with different modules devoted to both conceptual, linguistic, and graphical information.

In this paper, we will focus on some of the steps applied to extract and represent conceptual knowledge in EcoLexicon. According to Meyer et al. (1992), terminological knowledge bases (TKBs) should reflect conceptual structures in a similar way to how concepts relate in the human mind. The organization of semantic information in the brain should thus underlie any theoretical assumption concerning the retrieval and acquisition of specialized knowledge concepts as well as the design of specialized knowledge resources (Faber, 2010). In Section 2, we explain how knowledge is extracted through corpus analysis. In Section 3, we show how conceptual knowledge is represented and contextualized in dynamic and interrelated networks.

[1] http://ecolexicon.ugr.es

2. Conceptual Knowledge Extraction

According to corpus-based studies, when a term is studied in its linguistic context, information about its meaning and its use can be extracted (Meyer & Mackintosh, 1996). In EcoLexicon, the corpus consists of specialized (e.g. scientific journal articles, thesis, etc.), semi-specialized texts (textbooks, manuals, etc.) and texts for the general public, all in the multidisciplinary domain of the environment. Each language has a separate corpus and the knowledge is extracted bottom-up from each of the corpora. The underlying ontology is language independent and based on the knowledge extracted from all the corpora. The extraction of conceptual knowledge combines direct term searches and knowledge pattern (KP) analysis. According to many studies on the subject, KPs are considered one of the most reliable methods for knowledge extraction (Barrière, 2004). Normally, the most recurrent knowledge patterns (KPs) for each conceptual relation identified in previous research are used to find related term pairs (Auger & Barrière, 2008). Afterwards, these terms are used for direct term searches to find new KPs and relations. Therefore, the methodology consists of the cyclic repetition of both procedures.

When searching for the term EROSION, conceptual concordances show how different KPs convey different relations with other specialized concepts. The main relations are *caused_by*, *affects*, *has_location* and *has_result*, which highlight the procedural nature of the concept and the important role played by non-hierarchical relations.

In Figure 1, EROSION is related to its diverse kinds of agents, such as STORM SURGE (1, 7), WAVE ACTION (2, 13), RAIN (3), CONSTRUCTION PROJECTS (6) and HUMAN-INDUCED FACTORS (11).They can be retrieved thanks to all KPs expressing the relation *caused_by*, such as *resultant* (1), *agent for* (2, 3), *due to* (6, 7), and *responsible for* (11).

```
Caused_by

1    , Alabama. Significant storm surge and resultant beach erosion were  associated with Ivan's landfall. However,
2    nd climate  on the Castellón coast, the main agent for erosion is wave action, and this is therefore  responsi
3    f a stream. The first   factor, rain, is the agent for erosion, but the degree of erosion is governed by   oth
4    rts (BW) and semiarid steppe (BS). Wind can also cause erosion and deposition in environments where sediments
5    etty. Reflection of waves from a jetty may  also cause erosion of adjacent shorelines. However, erosion furthe
6    oastal zone management. However, in some cases coastal erosion\par  can be due to construction projects that a
7    tude of about 0,3 M m3 per year.  Acute erosion  Acute erosion due to storm surges  (waves and water levels at
8    er. Mangrove removal is also reported to cause coastal erosion and change sedimentation patterns and shoreline
9    [edit] Erosion  Surface runoff is one of the causes of erosion of the earth's surface. Reduced   crop producti
10   pes. Local disturbances, for instance by flood-induced erosion, redistribution of sediment or accumulation of
11   ors and human-induced factors  responsible for coastal erosion and highlight the time and space patterns withi
12   ocess is typical of a cyclical process of storm-caused erosion in winter, followed by progradation\par  owing
13   can cause excessive wave action that can lead to beach erosion. Trash dumped from boats can be washed up onto
14    that have reached base level develop broad valleys by erosion caused by meandering channels. The stream chann

Affects

15   ing these sensitive creatures.  In some cases, coastal erosion can have adverse effects on water quality and h
16   ine Depositional Coasts  The erosion of coastlines and erosion of sediments being carried to the   shoreline b
17    use of dredged material to restore beaches damaged by erosion. EPA works with the U.S. Coast Guard to regulat
18   reasonable points, though when push comes to shove and erosion threatens buildings, traditional beach maintena
19   ks and arches found on irregular rocky coastlines; and erosion provides the material which forms deltas and ba
20    near the base of the cliff. Constant undercutting and erosion causes the cliffs to retreat landward.

Has_location

21   ed by the position of sand accumulation\par  and beach erosion around littoral barriers. A coastal structure i
22   hes. Kuenen (1950) estimates\par  that beach and cliff erosion along all coasts of the world totals about 0.12g
23   ce and divergence of wave energy over an offshore bar, erosion downdrift of a structure such as a groin, sudde
24   proportional to the longshore transport rate, and\par  erosion takes place downdrift at about the same rate. T

Has_result

25    Excessive loads of silt and other sediments caused by erosion can suffocate bottom-dwelling plants and animal
26   islands or coral reefs.  Primary coasts are created by erosion (the wearing away of soil or rock), deposition
27   \par  transported. Beach material is also derived from erosion of the coastal formations caused by\par  waves
28   ed to the   passage of the ice. Shorelines produced by erosion of glacial till deposits differ   markedly from
29   beaches   and marshes, are being formed as a result of erosion   and transportation of unconsolidated material
30   ion of the seashore and a rise in SLR. The results  of erosion could lead to further seawater intrusion that c
31   fs are developed  in landslide debris. In this cliffs, erosion of softer material has created bays. The  expec
32   s of steep systems, a sea-level rise may cause coastal erosion resulting in profile steepening, and therefore
```

Figure 1: Non-hierarchical relations associated with EROSION

```
Is_a

33   vided by the  area (A) of the drainage basin      (1) Erosion is the natural process of removal of soil by wa
34   in the Netherlands, geomorphological processes such as erosion, transport and sedimentation of sandy materials
35   BURY AND DUXBURY, 1996).    Coastal processes such as erosion and accretion are site-specific, season-specifi
36   these  catchments include:  Stormwater impacts such as erosion, channelisation, sediment deposition and sedime

Type_of

37   eroded by shallow overland flow (sheet, rill and gully erosion) and delivered to the drainage network. Channel
38   m the great   local relief, the result of differential erosion by glacier ice. Figure 9-20   includes two sche
39   ing flood events, the dikes are subject to the lateral erosion of the river   trying to reoccupy its former co
40   d enlarges these small channels and generates headward erosion directed towards the aggrading active channel (
41   out five percent of the material on most beaches. Wave erosion of rocky coasts is usually slow, even where the
42    of the Earth's land surface is   dominated by fluvial erosion, lakes that do occur are threatened with either
43   ind   climate, topography and surface roughness.  Wind erosion risk applies only when soils are dry and not co
44   oportional to the steepness of the land surface. Water erosion is in proportion to the shear stress exerted by
45   ley to become both wider and deeper over time. Glacial erosion also results in a change in the valley's cross-
46   dominate in periglacial environments: nivation; eolian erosion and deposition; and fluvial erosion and deposit
47    erosion processes. 215 CHAPTER 13 EQUATIONS: SEDIMENT Erosion caused by rainfall and runoff is computed with
48   gineers to simulate  cross-shore beach, berm, and dune erosion produced by storm waves and water levels. The l
49   uctures constructed to date have resulted in shoreline erosion   in their lee. Furthermore, the key environmen
```

Figure 2: Hierarchical relations associated with EROSION

This relation can also be conveyed through compound names such as *flood-induced* (10) or *storm-caused* (12) and any expression containing *cause* as a verb or noun: *one of the causes of* (9), *cause* (4, 5, 8) and *caused by* (14). EROSION is also linked to the patients it *affects*, such as WATER (15), SEDIMENTS (16), and BEACHES (17). However, the affected entities, or patients, are often equivalent to locations (eg. if EROSION *affects* BEACHES it actually *takes place at* the BEACH). The difference lies in the kind of KPs linking the propositions. The *affects* relation is often reflected through the preposition *of* (10) or verbs like *threatens* (18), *damaged by* (17) or *provides* (19), whereas the *has_location* relation is conveyed through prepositions linked to directions (*around*, 21; *along*, 22; *downdrift*, 23) or spatial expressions such as *takes place* (24). In this way, EROSION appears linked to the following locations: LITTORAL BARRIERS (21), COASTS (22) and STRUCTURES (23). *Result* is an essential dimension in the description of any process, since it also has certain effects, which can be the creation of a new entity (SEDIMENTS, 25; MARSHES, 29; BAYS, 31) or the beginning of another process (SEAWATER INTRUSION, 31; PROFILE STEEPENING, 32).

All these related concepts are quite heterogeneous. They belong to different paradigms in terms of category membership or hierarchical range. For instance, some of the agents of EROSION are natural (WIND, WAVE ACTION) or artificial (JETTY, MANGROVE REMOVAL) and others are general concepts (STORM) or very specific (MEANDERING CHANNEL). This explains why knowledge extraction must still be performed manually, but it also illustrates one of the major problems in knowledge representation: multidimensionality (Rogers, 2004).

This is better exemplified in the concordances in Figure 2, since multidimensionality is most often codified in the *is_a* relation. In the scientific discourse community, concepts are not always described in the same way because they depend on perspective and subject-fields. For instance, EROSION is described as a natural process of REMOVAL (33), a GEOMORPHOLOGICAL PROCESS (34), a COASTAL PROCESS (35) or a STORMWATER IMPACT (36). The first two cases can be considered traditional ontological hyperonyms. The choice of any of them depends on the upper-level structure of the representational system and its level of abstraction. However, COASTAL PROCESS and STORMWATER IMPACT frame the concept in more concrete subject-fields and referential settings. The same applies to subtypes, where the multidimensional nature of EROSION is clearly shown. It can thus be classified according to the dimensions of *result* (SHEET, RILL, GULLY, 37; DIFFERENTIAL EROSION, 38), *direction* (LATERAL, 39; HEADWARD EROSION, 49), *agent* (WAVE, 41; WIND, 43) and *patient* (SEDIMENT, 47; DUNE, 48; SHORELINE EROSION, 49).

3. Dynamic Knowledge Representation

Since categorization is a dynamic context-dependent process, the representation and acquisition of specialized knowledge should certainly focus on contextual variation. Barsalou (2009: 1283) states that a concept produces a wide variety of situated conceptualizations in specific contexts. Accordingly, dynamism in the environmental domain comes from the effects of context on the way concepts are interrelated. Multidimensionality is commonly regarded as a way of enriching traditional static representations (León Araúz and Faber, 2010). However, in the environmental domain it has caused a great deal of information overload, which ends up jeopardizing knowledge acquisition. This is mainly caused by versatile concepts, such as WATER, which are usually top-level general concepts involved in a myriad of events.

Our claim is that any specialized domain contains sub-domains in which conceptual dimensions become more or less salient depending on the activation of specific contexts. As a result, a more believable representational system should account for re-conceptualization according to the situated nature of concepts. In EcoLexicon, this is done by dividing the global environmental specialized field in different contextual domains: HYDROLOGY, GEOLOGY, BIOLOGY, METEOROLOGY, CHEMISTRY, ENGINEERING, WATER TREATMENT, COASTAL PROCESSES and NAVIGATION.

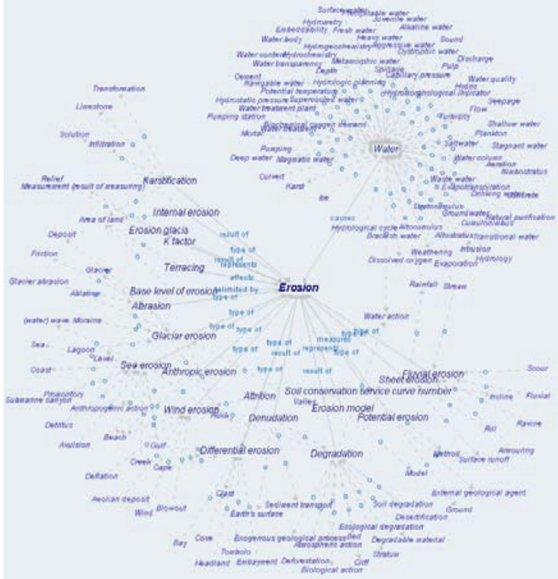

Figure 3: EROSION context free network

Nevertheless, not only versatile concepts, such as WATER, are constrained, since information overload can also affect any other concept that is somehow linked with versatile ones. For instance, Figure 3 shows EROSION in a context-free network, which appears overloaded mainly because it is strongly linked to WATER, since this is one of its most important agents.

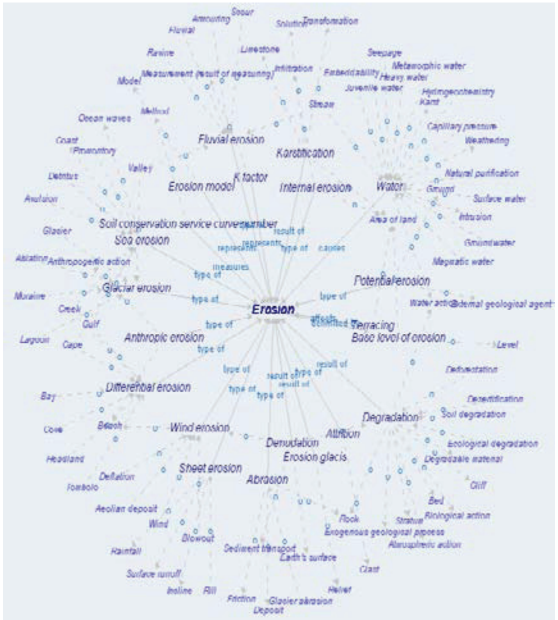

Figure 4: EROSION in the GEOLOGY domain

Contextual constraints are neither applied to individual concepts nor to individual relations, instead, they are applied to each conceptual proposition. When constraints are applied, EROSION is just linked to propositions belonging to the context of GEOLOGY (Figure 4) or

HYDROLOGY (Figure 5).

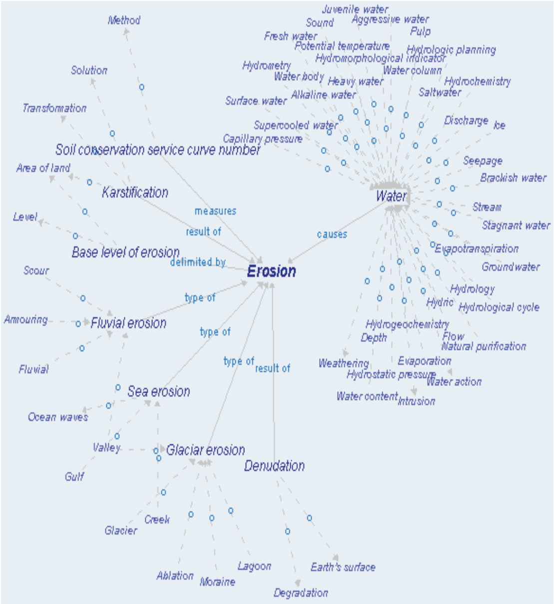

Figure 5: EROSION in the HYDROLOGY domain

Comparing both networks and especially focusing on EROSION and WATER, the following conclusions can be drawn. The number of conceptual relations changes from one network to another, as EROSION is not equally relevant in both domains. EROSION is a prototypical concept of the GEOLOGY domain, this is why it shows more propositions. Nevertheless, since it is also strongly linked with WATER, the HYDROLOGY domain is also essential in the representation of EROSION. Relation types do not substantially change from one network to the other, but the GEOLOGY domain shows a greater number of *type_of* relations. This is due to the fact that the HYDROLOGY domain only includes types of EROSION whose agent is WATER, such as FLUVIAL EROSION and GLACIER EROSION. The GEOLOGY domain includes those and others, such as WIND EROSION, SHEET EROSION, ANTHROPIC EROSION, etc. The GEOLOGY domain, on the other hand, also includes concepts that are not related to HYDROLOGY such as ATTRITION because there is no WATER involved.

On the contrary, WATER displays more relations in the HYDROLOGY domain. This is caused by the fact that WATER is a much more prototypical concept in HYDROLOGY. Therefore, its first hierarchical level shows more concepts. For example, in GEOLOGY, there are less WATER subtypes because the network only shows those that are related to the geological cycle (MAGMATIC WATER, METAMORPHIC WATER, etc.). In HYDROLOGY, there are more WATER subtypes related to the hydrological cycle itself (SURFACE WATER, GROUNDWATER, etc.). Even the shape of each network illustrates the prototypical effects of WATER or EROSION. In Figure 4, EROSION is displayed in a radial structure that shows it as a central concept in GEOLOGY, whereas in Figure 5, the asymmetric shape of the network implies that, more than EROSION, WATER is the prototypical concept of HYDROLOGY.

4. Acknowledgements

This research has been carried out in project FFI2011-22397/FILO funded by the Spanish Ministry of Science and Innovation.

5. References

Auger, A., Barrière, C. (2008): Pattern-based approaches to semantic relation extraction: A state-of-the-art. Special Issue on Pattern-Based Approaches to Semantic Relation Extraction, Terminology, 14(1), pp. 1–19

Barrière, C. (2004): Knowledge-rich contexts discovery. In Proceedings of the 17th Canadian Conference on Artificial Intelligence (AI'2004). May 17–19, London, Ontario, Canada.

Barsalou, L.W. (2009): Simulation, situated conceptualization and prediction. Philosophical Transactions of the Royal Society of London: Biological Sciences, 364, pp. 1281–1289.

Faber, P. (2010): Conceptual modelling in specialized knowledge resources. In XII International Conference Cognitive Modelling in Linguistics. September, Dubrovnik.

León Araúz, P., Faber, P. (2010): Natural and contextual constraints for domain-specific relations. In Proceedings of Semantic relations. Theory and Applications. 18–21 May, Valetta, Malta.

Meyer, I., Mackintosh, K. (1996): The corpus from a terminographer's viewpoint. International Journal of Corpus Linguistics, 1(2), pp. 257–285.

Meyer, I., Bowker, L., Eck, K. (1992): COGNITERM: An experiment in building a knowledge-based term bank. In Proceedings of Euralex '92, pp. 159–172.

Rogers, M. (2004): Multidimensionality in concepts systems: a bilingual textual perspective. Terminology, 10(2), pp. 215–240.