Knowledge Extraction and Multidimensionality in the Environmental Domain[1]

Pilar León Araúz and Arianne Reimerink

Keywords: knowledge extraction, multidimensionality, corpus pattern analysis, knowledge pattern

1 Introduction

The systematization of knowledge extraction techniques would make knowledge representation a more empirical and efficient process. But first of all, certain selection criteria must be defined by manually identifying what kind of information is useful or not, why it is so and how it is structured, so that automatic processes can be improved with human implicit encyclopaedic knowledge. Knowledge patterns (KPs) have been widely explored and are considered as one of the most reliable methods in knowledge extraction (Cimiano and Staab 2005; Marshman and L'Homme 2006). However, many of them are domain-specific and cannot always be reused. Moreover, they often convey different conceptual relations and are therefore polysemic structures. The development of pattern-based constraints can help to disambiguate them and at the same time avoid conceptual noise. Nevertheless, another factor that must be taken into account in knowledge representation is contextual multidimensionality. A natural consequence of multidimensionality is conceptual dynamism, especially prevalent in process-oriented domains, such as the ENVIRONMENT[2]. Even within the same specialized domain, concepts are involved in very different frames or situations and, consequently, they may experience a change in their relational behaviour that must somehow be represented. In the following sections we show how Corpus Pattern Analysis can tackle both dynamism and knowledge patterns in conceptual concordances.

2 Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) (Pustejovsky, Hanks and Rumshisky 2004; Hanks and Pustejovsky 2005) investigates syntagmatic criteria for distinguishing different meanings of a polysemous predicate. The procedure consists of three components: (1) the manual discovery of selection context patterns for specific verbs; (2) the automatic recognition of instances of

[1] This research has been supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation.

[2] Our research on environment is hosted in http://manila.ugr.es/visual/

the identified patterns; (3) the automatic acquisition of patterns for unanalyzed cases (Rumshisky *et al*. 2006: 329).

We apply the CPA approach in a slightly different way. On the one hand, the extraction of concordances starts with a direct search of specialized terms. After that, they are classified according to the dimensions they show (*material*, *function*, *location*, etc.), where different knowledge patterns can be associated to different conceptual relations (*made_of*, *has_function*, *located_at*, etc.). Then we analyse multidimensionality according to context, since conceptual dimensions are sometimes restricted to specific situations.

In this sense, Rumshisky *et al.* (2006) indicate several ways in which different context dimensions expressed in the selection context patterns can affect the semantic interpretation of a predicate. According to them, *the most frequent source of meaning differentiation of verbs lies in contrasting the argument types filling each argument slot (*Rumshisky *et al*. 2006:330). However, in our research CPA is applied to identify how context can affect the relational behaviour of concepts, mainly noun entities. In our experience, *the most frequent source of context differentiation of a concept's behaviour lies in contrasting the specific values filling each dimension*. This does not mean that concepts are regarded as polysemic structures (this could only be the case of terms), but rather that the activation of different contexts makes certain conceptual relations (or dimension values) incompatible. In Tables 1 and 2 both approaches are shown.

Table 1: Concordances of *abandon* in three different senses according to the CPA approach[3]

Meaning differentiation of the verb *abandon*
[[Human | Institution]] abandon [[Activity | Plan]] [[Human | Institution]] *stops doing* [[Activity]] *or does not begin to do* [[Plan]] **Concordances** did not agree that Britain could or should **abandon** development, either for itself or for the After discreet soundings, they prudently **abandoned** the idea, which would have involved a
[[Human | Institution]] abandon [[Attitude]] [[Human | Institution]] *ceases to have* [[Attitude]] **Concordances** had previously attacked, and in 1931 he **abandoned** the principle of free trade, which he had s hope for a future society, which would **abandon** the obsession with private property, but
[[Human 1 | Animal 1]] abandon [[Human 2 | Animal 2]] (to [[Anything = Bad]]) [[Human 1 | Animal 1]] *goes away from and ceases to care for or look after* [[Human 2 | Animal 2]], *with the result that* [[Anything=Bad]] *may get them or happen to them* **Concordances**

[3] http://deb.fi.muni.cz/pdev/?action=patterns&id=abandon

to join sects that make their adherents **abandon** their families, give up their money and t posed by thousands of animals being **abandoned** or buried in the countryside, according

Table 2: Concordances of WATER in three different domains according to our corpus analysis approach

Context differentiation of the concept WATER
AGRICULTURE [water + (functional pattern: *available for…*) + agriculture-related concept] supplies in the area. In areas where some **water** is *available for* irrigation shrubs planted on
WATER TREATMENT/SUPPLY [water + (functional pattern: *for use by…*) + water treatment/supply-related concept] accompanying increased demand of fresh **water** *for use by* a growing population. One of the
GEOLOGY [water + (causative pattern: *caused…*) + geology-related concept] s vegetation to hold soil in place, flowing **water** *caused* even more erosion and nutrient loss, which ley has been limited and linear erosion *by* **water** has been a major geomorphic process in the all

According to our corpus analysis approach for noun specialized concepts, different collocational values activate diverse situations and referents that do not usually coincide, if ever, in time or space. In Table 2, WATER is still the same concept in all three examples, but the conceptual propositions in which it participates are context-dependent. In a more natural context, WATER is usually regarded as an agent, which explains why it is prototypically linked through causative patterns. However, in anthropic contexts, WATER is more generally regarded as an artifact or a patient, which is often illustrated through functional patterns.

On the other hand, a pattern-based search is also conducted in order to find new relations among other concepts. Since many KPs are codified by polysemic verbs, a more classical CPA approach can be applied to search for a "real" meaning differentiation. For instance, it is evident that *form* does not have the same meaning in the following sentences: clouds are *formed of* water and clouds are *formed by* condensation. In this case prepositions play an important role, but also do argument fillers. The second arguments of each sentence (*water* and *condensation*) have a different conceptual nature (entity and process) and imply different relations (*made_of* and *result_of*). Our approach differs in the type of features we use to disambiguate them. Whereas in classical CPA arguments are all classified as semantic categories, we also add other types of information, such as morpho-syntactic data. However, not all verbs used as KPs are polysemic structures, such as *range* or *constitute*. In these cases, only argument fillers and their specificity inform us about context, dimensions and relation types, which also avoid conceptual noise.

3 Conceptual dimensions

In the first step, apart from manually extracting all dimensions related to candidate terms, we found that conceptual dimensions do not have the same status or relevance, and they have different implications depending on their representational level and specific contexts. We propose a three-level classification: prototypical dimensions, additional dimensions and dynamic dimensions (León Araúz 2009).

3.1 Static knowledge: prototypical and additional dimensions

Prototypical dimensions are the most frequent dimensions found in the corpus. They match concepts' definitional structure, which is also a valuable source for top-down knowledge extraction. As a result, their values describe concepts in a static context, which means that they are intrinsic properties that should be modelled in any kind of representation. For instance, the *differentiae* observed in several definitions of GROIN correspond to *material*, *location* and *function* dimensions.

Table 3: Conceptual dimensions found in definitions of GROIN

Concept 1	Dimension	Concept 2
Groin	***Material***	wood; stone, concrete; timber, rock; sheet-pile, rubble-mound
	Location	coastline; shore; shore; seawards; shore, water; sea, river; shore; seawall, backshore, foreshore; backshore, water body
	Function	prevent erosion and longshore drift; control erosion; stop the flow of beach material moved by longshore drift; prevent a beach from washing away; protect the shore from erosion; reduce longshore currents; trap littoral drift, retard erosion of the shore

Most concordances showed the same information, through different KPs like *made of*, *constructed of*, *built out from*, *located along* and *function by*.

Figure 1: Knowledge patterns and conceptual dimensions in concordances of GROIN

```
MATERIAL

sand, nourishing the beach  compartments between them. Groins may be made of wooden or rocky materials. They
s designed as rubble-mound structures (breakwaters and groins) except that the outer part must be armored on b
 variety of materials  are used in the construction of groins. Impermeable groins can be constructed  of stone

LOCATION

olution includes the following  three elements: a rock groin extending seaward  from the existing shoreline, a
are intended consequences for people and wildlife.     Groins are structures built out from the shoreline, typ
San Luis Pass (Morton 1993). There are 15 rubble stone groins located along the beach between 10th and 61st St

FUNCTION

ions to anchor the fill material. In either instance,  groins provide shore protection by modifying longshore
 approach to controlling beach erosion. However, since groins function by trapping sand within the littoral sy
nsions or natural beach elevation) trap sand like high groins. However,  some of the sand also passes over the
ngth of beach to be protected. The basic purposes of a groin are to modify the longshore movement of sand and
```

However, concepts can also be related in many other ways in the real world. Additional dimensions add new knowledge in two ways: (1) new conceptual dimensions may arise, (2) or conceptual dimensions remain the same, but they show different values from those found in the definition, some of them having a different degree of focalization compared to definitional ones, and some others belonging to diverse hierarchical paradigms.

Figure 2: Additional dimensions of GROIN

```
PART

irst  in year 1983) and 2 submerged groins linking the groin head to the barrier (the northern built in 1995 a
erosional depression in beach form, as in the lee of a groin, moves downdrift as a traveling sand deficit beca

MATERIAL

groins. In addition to the above  materials, permeable groins can be made of sand bags, large stones, and eart

LOCATION

eable or impermeable, and fixed  or adjustable. A high groin, extending through the surf zone for ordinary or
ters to flow underneath their structures.  Sand Traps  Groins and jetties — structures built perpendicular to

CAUSE

raphy changes   As can be seen from Figs. 6 and 7, one groin caused remarkable accretion at the updrift   (wes
ds. Shortterm  events of elevated turbidity induced by groin construction or associated  beach fill will occur

FUNCTION
 accumulation. In some cases may serve as a "terminal" groin to anchor a nourished beach.  Groin System or Fie
 performed a detailed study to assess the effects of a groin and beach fill improvement on recreational surfin
```

In Figure 2, the *material*, *location* and *function* dimensions are still associated to GROIN. Nevertheless, GROIN appears related to more specific concepts: SAND BAGS, SURF ZONE, ANCHOR, etc. In addition, new dimensions like *part* and *cause* widen its relational power with new concepts like HEAD, LEE, ACCRETION and TURBIDITY. Additional dimensions are less relevant than prototypical dimensions and their inclusion in a knowledge base depends on its level of specialization.

The extraction of both prototypical and additional dimensions only relies on the reliability of KPs, which will be dealt with in section 4. However, dynamic dimensions need a CPA approach, since any specialized domain contains sub-domains in which conceptual dimensions become more or less salient depending on the activation of specific contexts.

3.2 Dynamic dimensions: contextual knowledge

Dynamic dimensions reflect a contextual multidimensionality derived from the situated nature of concepts. They are determined by context and occur especially in the case of concepts with a low degree of specificity. We call them *versatile concepts* because they are involved in a myriad of events. However, even if they appear in different contexts, they are not always related to the same concepts or through the same relations, especially in such a wide domain as the environment, where entities are subject to constant transformation.

For example, the concept SAND is generally (or prototypically) defined as a kind of sediment located in the sea, rivers or soil layers. A specialized definition could even include some information about its grain size. However, in real texts, SAND activates many other relations. In a more general domain, such as GEOLOGY, the concept is linked to others through: *type*, as a kind of SEDIMENT; *attribute*, related to grain size as a classification parameter; and *material*, linking the concept to the natural elements of which it forms part (VALLEY, SOIL, AQUIFER, DESERT, etc).

Figure 3: SAND in the GEOLOGY domain

```
GEOLOGY

TYPE

, suspended sediment has been categorized as suspended sand, silt, and   clay. The transport of each class is
ot drain by gravity and earthen material such as rock, sand, gravel,   or clay. This means that a 10 unit drop

MATERIAL

consisting   of a heterogeneous mixture of clay, silt, sand, gravel, and boulders ranging   widely in size and
n Maine. Sediments are composed of fine clays,   silt, sand and organic matter. Sediments are supplied to the
s have been proposed to classify sediments composed of sand-silt-clay mixtures in natural systems without, how
arbonate) areas or where soils and aquifers consist of sand and gravel. These natural features enable rapid in

ATTRIBUTE

 Petri parallel at  a depth of 200 m in muddy and fine sand sediments,  representing 2 % of the total assembla
 A comprises sandy sediments sites (medium  and coarse sand) with a low percentage of fines  and total volatil
 decelerated, sediments ranging from medium boulder to sand-sized were deposited. Sedimentation was controlled
```

These dimensions could be considered additional ones. However, some of them are not compatible in other contexts. In a COASTAL PROCESS domain, salient dimensions become: *material*, although natural elements are restricted to coastal ones (SAND BARRIER, SAND BERM, SAND SPIT, BEACH, etc.); and *patient*, where the concept is involved in certain natural processes (WAVE ACTION, STORMS, LONGSHORE CURRENT, DEPOSITION, etc.).

Figure 4: SAND in the COASTAL PROCESS domain

```
COASTAL PROCESS

MATERIAL

  A beach is an area of sediment accumulation (usually sand) exposed to wave   action along the coast. Beache
 was designed along the entire  stretch of SR-105 as a sand berm with crest elevation  at +30 feet MLLW. Sand
sion and producing broader beaches.  Spits and bars  A sand spit is one of the most common coastal landforms.
e  sedimentary terraces or "raised beaches" containing sand and gravel deposits. The coastline  forms a chain
lluvium (debris from valley sides),   channel deposits (sand and gravel), and vertical accretion deposits (cla
f medium to coarse sand. Channel banks are composed of sand and  mud (silt-clay content averaging 30-40%) and

PATIENT

by sand   bars, which are formed by wave action moving sand onto and along the beach. The   river is then only
 level factors.\par  Most storms move large amounts of sand from the beach to offshore; but after the storm, t
ngle. The longshore current can carry large amounts of sand along the coast and can form spits (narrow peninsu
  longshore drift. Longshore drift erodes and deposits sand continuously along the   beach. The sand that is r
 different   areas of the continental shelf and slope. Sand being the largest, is transported   by waves towar
 action. During the summer months when waves are low,  sand is deposited on the beach, forming a high and wide
```

If context is again restricted to the COASTAL DEFENCE domain, dimensions are still the same, but values are focalized to artificial elements (FENCE, BERM, DIKE) or processes (TRAPPING, PUMPING, DUMPING). Furthermore, there is a new dimension, highlighting the functional nature of the concept in this specific context (SAND *protects* DUNE-BLUFFS, SAND BODY *is used for* BEACH NOURISHMENT, etc.).

Figure 5: SAND in the COASTAL DEFENCE domain

```
COASTAL DEFENCE

MATERIAL

 them recover   naturally. (See also Case 2.) However, sand fences can be erected to help dune recovery after
ss impacts associated with construction of a nearshore sand berm. Baseline condition descriptions included dis
andward 550 ft of the west jetty were constructed as a sand-fill dike, with a crown elevation of +4 ft mlt and

FUNCTION

e also important to many beaches because they act as a sand   source. Many fish actively feed on the coral. Fo
important beneficial effects. First, beach nourishment sand directly protects the natural dune-bluffs from wav
 sources of sand.  Sand Sources for Beach Nourishment  Sand for nourishment projects is from a variety of envi
icate a highly variable environment. Thus, before this sand body is used for beach nourishment, further coring

PATIENT

der to restore it to its former width. The addition of sand  to the beach by dredging and pumping sand from of
ng wave energy in the case of breakwaters and trapping sand in the case of groins, thus influencing the sand m
jacent beaches, artificial sand bypassing can be used. Sand bypassing is the hydraulic or mechanical movement
movement of sand generally can  be restored by pumping sand from the side where sand accumulates through a  pi
urishment for the\par  beach.\par  Offshore dumping of sand by hopper dredge was carried out at Long Branch, N
```

These three domains form a hierarchy (GEOLOGY → COASTAL PROCESS → COASTAL DEFENCE), but in a completely different domain, changes are more remarkable.

Figure 6: SAND in the WATER TREATMENT domain

```
WATER TREATMENT

INSTRUMENT

e water column. This equipment is called a detritor or sand  catcher. Sand grit and stones need to be removed
 5 1 of water were collected after filtration  through sand filters. Chlorine concentration used in  the exper
d grit removal  Primary treatment typically includes a sand or grit channel or chamber where the   velocity of
first nitrifying the wastewater by passage through the sand filter, then recirculating the nitrified effluent

FUNCTION

lso called   "effluent polishing".  [edit] Filtration  Sand filtration removes much of the residual suspended
water-treatment plants in the Kingdom utilize imported sand for filtration. The objectives of this research pr
beds for wastewater sludge require a specific type of sand in order to dewater the sludge quickly. Author War
```

In the WATER TREATMENT domain, a new dimension is found, where SAND is linked to a particular instrument used in water treatment plants. The functional dimension now has a different value (FILTRATION) and *patient* and *material* are no longer representative conceptual dimensions.

Context is a dynamic construct that triggers or restricts knowledge (León Araúz *et al.* 2009). Moreover, Yeh and Barsalou (2006) claim that when situations are incorporated into a cognitive task, processing becomes more tractable than when situations are ignored, and the same can be applied to knowledge acquisition processes. As a result, a more believable representational system should account for re-conceptualization according to the situated nature of concepts, which would enhance knowledge acquisition.

In our research, we are developing context-based conceptual networks by dividing the environmental field in different contextual domains. Nevertheless, first we need to know *which* concepts are activated in each situation and *how* to extract this information. This is where CPA can help to accomplish our aim.

According to CPA, patterns expressing contextual information delimit the lexical senses of words. In our approach, patterns expressing contextual information delimit conceptual dimensions, which at the same time delimit domain membership.

Two procedures must be applied. First of all, if a concept only activates a particular dimension in just one domain, the identification of any KP linking the concept to that dimension will be enough to associate the concept to a concrete domain. In the case of SAND, if a KP expresses the *instrument* dimension, the concept will be automatically assigned to the WATER TREATMENT domain.

Nevertheless, *domain disambiguation* is not that easy when different domains can activate the same dimensions and, as a result, KPs are usually the same. For example, if SAND is found next to KPs like *consist of, comprising, formed from, containing,* or *composed of*, the *material* dimension ascribes concepts to three possible domains: GEOLOGY, COASTAL PROCESSES and COASTAL DEFENCE. Sometimes, certain KPs express a particular dimension in just one domain, such as *made of*, where SAND is always related to COASTAL DEFENCE because the pattern needs the activation of an artificial concept.

However, most of the time, this is not the case, and domain disambiguation requires a second step based on the kind of values associated to each dimension. At this stage, semantic annotation seems the only way to differentiate domain membership. As mentioned above, in the GEOLOGY domain materials must be natural elements found in nature, in the COASTAL PROCESS domain, materials are restricted to those found in the coastal area, and in the COASTAL DEFENCE domain materials are no longer natural elements. Consequently, annotation should be concept-oriented, differentiating all concepts in the hierarchy and assigning each of them to particular contexts. In this way, when dimensions are common to several domains, dimensional values are the only thing that can guide the knowledge extraction process. Semantic annotation has not been implemented yet, since first of all we need to know which kind of information (e.g. conceptual, semantic, morpho-syntactic) must be included, which is also the case for KPs.

4 Knowledge patterns

Many KPs can be found in a manual identification process. However, automatic extraction needs a certain level of reliability to be effective. In table 2 we show some of the most reliable patterns for seven conceptual relations in our specialized domain.

Table 3: Knowledge patterns and their conceptual relations

Conceptual relation	**Knowledge pattern**
Is_a	such as, rang* from, includ*
Part_of	includ*, consist* of, formed by/of
Made_of	consist* of, built of/from, constructed of, formed by/of/from
Located_at	form* in/at/on, found in/at/on, tak* place in/at, located in/at
Result_of	caused by, leading to, derived from, formed when/by/from
Has_function	designed for/to, built to/for, purpose is to, used to/for
Effected_by	carried out with, by using

We have applied CPA to certain general patterns. For example, despite all the information that *form** helps to extract, it would have been a non-reliable pattern if no prepositions or verb tenses were included as part of the pattern. Thus, when *form** is followed by *in*, *at* or *on*, it only expresses the relation *located_at*.

Domain-specific patterns are generally more reliable, such as *built of/from* and *constructed of*, which have 100 per cent reliability. But most of them are general language patterns that can be applied to many other domains. On the other hand, even reliable ones show a certain degree of conceptual noise. A way of solving this problem could be to search for candidate terms of interest occurring with patterns indicating relations that may be useful for describing the concepts they denote, but that would rule out unidentified terms.

Semantic annotation seems again a plausible way to exclude useless information. For example, in the KP *consist* of*, a higher level of reliability would be achieved if automatic searches could be restricted to part of speech *nouns*, since parts or materials will never be anything other than nouns. In this way, sentences like "the process *consists of* building a groyne 97m long" would be omitted. On the other hand, if concept types were also annotated in the corpus, it would be easier to deal with KPs like *carried out with*. In this case, only concepts sharing the category membership of instrument would be selected, avoiding sentences like "two test series were *carried out with* the following parameters".

This can also be applied to KPs expressing several conceptual relations, as in the case of domain disambiguation. Certain criteria could be elaborated for patterns like *includ** (*type_of* and *part_of*), *consist of* (*part_of* and *made_of*) or *formed by* (*part_of*, *made_of* and *result_of*). Nevertheless, before annotating any corpus, it is necessary to discover certain rule constraints according to each KP's specific needs, whether they cause conceptual noise or they are polysemic structures. In the next sections we will deal with the KPs *rang* from* and *formed by*.

3.1 KP rang from*

One of the most reliable knowledge patterns that activates the relation *is_a* in the environmental domain is *rang* from*. In our corpus there are 1381 occurrences of this pattern. Still, noise appears when studying the concordances with this knowledge pattern. When *rang* from* is followed by a number (see Figure 7), the relation that is expressed is always of magnitude. This is interesting information for the description of prototypical and additional dimensions of some types of concepts, but not if our aim is to analyze the conceptual relation IS_A. If we omit all these concordances, there are 356 concordances left. The same goes for the combination with a number written in full, or an adverb or an article and a number. Measurable concepts such as GROWTH TIMES and RECHARGE EVENTS are defined by certain time and amount combinations, which are additional dimensions. Therefore, if our aim is to analyze the IS_A relation, there should also be restrictions on words expressing duration such as minute, hour, month, day, week, etc. (see Figure 8).

Figure 7: Noise in the KP *rang* from*: numbers

```
ditions. Temperatures in Italian rice fields typically range from 15 to 30 øC (Sch□tz et al., 1990). We chose
ved at different locations with aggregation timescales ranging from 15 min to 2 days. The theoretical and obse
 a marked thermocline  around 30 m depth; temperatures ranged from 15 to  17Â°C in the epilimnion and 7Â°C in
by land clearing,  grubbing, grading, and construction range from 149,459  to 198,494 acres, or 25 to 33 perce
he same image analysis  system as in the present study ranged from 15 to 90 mg  C mâ€“3 from March to October
29.5 km upstream of the reservoir. Elevations of sites range   from 1470 to 1700 m MSL (Fig. 4). To facilitate
ese systems have a projected daily flow of wastewater  ranging from 15,000 to 4,000,000 gpd. All permitted lan
V, between Sines  and Cape St Vicente, at water depths ranging from 15-387 m.  During these cruises, equinoder
as calculated for copepods by Frost (1972), and values ranging from 15.000 to 50.000 cells ind-1 h-1 were repo
1). The climate is semi-arid with mean annual rainfall ranging from 15 to 28 cm, increasing to the southeast.
turns to the west across faults with sinistral offsets ranging from 15 m to 100 m (Fig. 3B). Between the Wuton
```

Figure 8: Noise in the KP *rang* from*: magnitude

```
eynolds number (Rec), based on the collector diameter, ranges from well below 1 to 1,000. No analytical soluti
ow  downgradient wells. Water table depth in the wells ranged from approximately 5 to 8 feet below the  ground
eas that are  stratified into 79 beach sampling units, ranging from  ca. 3 to 37 km in length, and which are m
9]). Throughout the collected samples in 1999, the TDS ranged from a minimum of 5900 mg/l to a maximum of 14,2
 cannot proceed much because the bedrock is at a depth ranging from about 100 to 120 m. However, several secto
ater aquifers  of North Carolina, pesticide detections ranged from  less than 1 to 34 percent of standards or
tandards or guidance  levels. The remaining detections ranged from less  than 1 to 50 percent of standards or
n plane and relatively steep beaches with growth times ranging from hours to one month. Less steep equilibrium
e intervals for observable ET and rain recharge events range from days in the forest to months or years in the
 or average   state of the atmosphere over time scales ranging from decades to millions of   years. These chan
d for pesticides with  leaching potential ratings that ranged from very low to very high. The soil series reco
ns, average grain speeds on the Iron Mountain piedmont range from   a few decimeters to a meter per year. Such
ies are normally graded gravel-sand couplets (Fig. 6), ranging from decimeters to meters in thickness, occasio
ce structures corresponds to   solution-induced basins ranging from hundreds to several thousands of meters in
```

If these concordances are omitted, there are only 121 concordances left. The knowledge pattern *rang* from* has proven to be a very reliable pattern if the above-mentioned restrictions are taken into account. It is not only very informative for the extraction of hyponymic relations, but also for relations of coordination, as the expression requires at least two coordinate concepts.

Figure 9: Hyponymic and coordination concordances of the KP *rang* from*

```
from the natural beach profile must be built. Seawalls range from vertical face structures such as massive gra
graded, with uneven erosional bases. The gravel clasts range from small pebbles to large boulders, are subangu
ecord sedimentation in shallow playa-lake environments ranging from subaqueous to vadose and subaerial zones.
f 3 days.  Sediments  Sediments of different textures, ranging from  mud to sand, were collected from the inte
he  Clyde Sea area (Fig. 1). The sediments at 4 sites  ranged from mud to fine sand and all amoebae were  isol
s, the river was braided with individual anabranches   ranging from shallow, wide channels to deeper, narrow c
e   processes to attack the bedrock and produce debris ranging from silt-sized   particles to large boulders.
omea, 1997). Parameterizations of mixed layer dynamics range from relatively simple bulk models ( Kraus and Tu
and Australia. The types of juvenile           habitat range from the entire estuary or shallow open coastal w
ought of as being composed of a mosaic of slope types, ranging from steep mountains and cliffs to almost flat
sheries 1998).  Marine waters   The marine environment ranges from the intertidal zone to the vast waters of t
ia testudinum). Seagrasses normally occur in sediments ranging from  sand to mud in relatively protected envir
 occur in both freshwater and marine environments that range from pelagic to benthic and littoral to deep-sea
 number of species  with different feeding strategies, ranging from phagotrophic  ingestion of particles to ph
ow-to-moderate wave energy-coasts with beach sediments ranging from fine sand to pebbles [Dally and Pope, 1986
rada (Tarragona) is composed of different beach-types, ranging from long,   straight beaches to pocket beaches
er.         Where a variety of coarse material exists, ranging from pebbles to large         boulders, one fin
 three-dimensional computer pictures of ocean animals, ranging from plankton to large fish. This new technolog
ciomarine sediments   (which is a mixture of sediment, ranging from clay to   boulders). These were later rewo
stewater nitrogen signatures in a number of locations, ranging from estuaries to freshwater ponds, to a salt p
e   processes to attack the bedrock and produce debris ranging from silt-sized   particles to large boulders.
stry of the magma of these volcanoes is quite variable ranging from basalt to granite. Magmas that are more gr
 enclose a widespread record of sedimentary deposits   ranging from (i) the "classical" freshwater model in wh
f open-work gravels, but generally a variable  matrix, ranging from silt to very coarse sand, fills the voids.
```

From the concordances of figure 9, the following information can be extracted: FINE SAND and PEBBLE *is_a* BEACH SEDIMENTS; LONG BEACH and STRAIGHT BEACH *is_a* BEACH; SAND, MUD, CLAY and BOULDER *is_a* SEDIMENT; PHAGOTROPHIC INGESTION *is_a* FEEDING STRATEGY.

3.2 KP *formed by*

As in the case of *rang* from*, certain constraints must also be defined for the KP *formed by* in order to avoid noise. First of all, concordances where the KP is followed by a figure or a time expression (*formed by then*, *formed by 2.5 billion years ago*) must be omitted.

Figure 10: Noise in KP *formed by*: figures and time

```
  about 70 percent of the Earth's continental crust was formed by 2.5 billion years ago. Over the next 2 billio
 complete set of orthogonal equations over the depth is formed by (2.5), (2.6) and (2.7). To gain the full solu
 be almost 90ø, because the second vortex will not have formed by then. The vortex pairing will occur at near 9
```

Moreover, it is also possible that the KP links a specialized concept with a too general one, both to the right and the left of the pattern.

Figure 11: Noise in KP *formed by*: too general concepts to the left

```
 on of the state. It is  a watershed-based organization formed by a prescribed legal  process in Superior Court
 umber of setae (ca. 68) in the posterolateral complex  formed by these adjacent groups. In other decapod speci
```

Figure 12: Noise in KP *formed by*: too general concept to the right

```
 al plain  EUROSION Case Study  3  The examined area is formed by the interaction between river-delta and marin
 ogdan & Gilbert 1982).  The third functional group was formed by the algivorous  ciliates Coleps and Prorodon,
 period is the unknown, the\par  exact solution must be formed by an iterative procedure (or use of the tables
 section of Segment A. It is unrealistic that Segment A formed by rock fall from bedrock cliffs after the forma
 en H)R. A geophysical equivalent of  this situation is formed by leads, which have been modelled experimentall
 ohesive to cohesive, and (ii) the network structure is formed by another sediment fraction. Appropriate parame
   refuse to grow. An interesting exception to this is  formed by the isolates obtained of Rehnstam et al.  (19
 ng and by spit extension, although most appear to have formed by spit extension in the classic manner describe
 nean  Sea, and Skagerrak. A separate cluster was  also formed by 3 different Hyphomonas species from  the Skag
 4, ? = 0.76  (right side). The fifth obstacle (“E”) is formed by two hills of shape “A” joined up with a  cent
```

Finally, a significant problem common to all KPs, as in *rang* from*, is the fact that the first term is so far from the pattern, that it cannot even be manually extracted.

Figure 13: Noise in KP *formed by*: anaphora

```
red spot' (these were not counted because they can be  formed by bursting during the preparation of the  slide
lar layers   of compact and light coloured appearance, formed by massive conglomerates and   clays, strongly i
s reaching to about mean sea\par  level that have been formed by calcium carbonate-secreting organisms. The mo
r is the simplest case of an inhomogeneous flow. It is formed by two  layers of fluids having different densit
8    . or Jiang and Garwood 1996; 1998 , but must have formed by another mechanism, as  the density current wa
lated sand along the outer part of the shore. They are formed by the   motion of waves and tides.   When a lar
refore equally impossible        to predict. They are formed by a sudden uplift or subsidence of the sea
s are rounded and about 0.75 mm in diameter. They were formed by a nucleus, made of a detrital grain or root d
```

Noise shown in figures 11, 12 and 13 could be partly solved if a candidate term extractor could be implemented at the same time. This way, terms which did not fall under the consideration of specialized terms would be excluded from the extraction process from the beginning. Anaphora would still be a problem, but the validation of conceptual propositions would be at least more efficient. However, the main problem of the KP *formed by* is polysemy. Concordances in figure 14 show the way *formed by* works in the three different dimensions it can express, although the *result* dimension prevails over *part* and *material.*

Figure 14: Conceptual dimensions expressed by KP *formed by*

```
RESULT

 is the disturbing force of a seiche. Seiche is a wave formed by the rocking of water in an enclosed water are
s valley is a topographic low about 15 km wide and was formed by shearing along the transform that separates t
e lower berm is the natural or\par  normal berm and is formed by the uprush of normal wave action during the o
urce: Geomophology From Space   Delta coasts are those formed by the deposition of sediment at the mouth of a
effects). Also horizontal heterogenity such as burrows formed by biological activity, will affect the erosion
ep-walled ~80-m deep gorge in rock avalanche   debris, formed by headward erosion of seepage-fed streams emerg
s lie adjacent to ledge outcrops. Boulder beaches were formed by glacial scouring   and deposition and bedrock
level fluctuations. The higher berm, or storm berm, is formed by wave action during storm\par  conditions. Dur
 sheets by waning flood flows. These sandstones can be formed by either direct fallout sedimentation (massive
he rest of the year. Consequently,   coastal sediments formed by evaporation of seawater are common, particula
 a historic phenomenon; the world's great deserts were formed by natural processes interacting over long inter
 Auldyn alluvial fans (Fig. 1). The alluvial fans were formed by the deposition of vast quantities of sediment
Deep-water waves: Sea and swell   Wind waves initially formed by the action of wind blowing over the sea surfa
 a sharp jolt or shaking. P-waves or primary waves are formed by the alternate expansion and contraction of be
non-marine processes, and secondary coasts, which were formed by marine action. Primary coasts happen because

PART

1989) and form ice. Barents Sea bottom water (BSBW) is formed by cold, hyper-saline cascades, which slide down
olidated rocks . The Earth's surface in most places is formed by soil and by unconsolidated deposits that rang
ohesive to cohesive, and (ii) the network structure is formed by another sediment fraction. Appropriate parame

MATERIAL

 et al., 1999]. Both units correspond to alluvial fans formed by breccia and conglomerates, which distally pas
r  the Roman community.    The  cuspated  delta  was  formed  by  alluvial  sedi- ments  carried  by  the  ri
s are usually dark and regular with different laminae, formed by dark micrite with few detrital grains and by
```

The disambiguation of this polysemic KP requires different steps. If the KP is followed by a verb, it is definitely related to the *result* dimension. Instead, if the KP is followed by a noun, it can be linked to any of the three dimensions. Then the difference lies in two factors: if the noun is a process concept type, the concept still falls into the *result* dimension; if the noun is an object concept type, the dimension can be either *part* or *material*, but if the noun is uncountable, it will always refer to the *material* dimension, whereas countable nouns will always link wholes with *parts*.

In these concordances the word span among concepts and patterns is still quite wide, although there are no cases of anaphora. For *formed by*, a candidate term extraction process should also be running at the same time, where a 10 word maximum span would be set on the left side of the KP and a 4 word maximum span on the right.

5 Conclusions and future work

In order to construct a knowledge resource based on the real behaviour of concepts in all the possible contexts of a domain, all the dimensions of multidimensional concepts must be taken into account. Therefore, apart from providing a definition which conveys their prototypical features, additional and dynamic dimensions must also be made explicit. In our research, we use the different dimensions of concepts to restrict the conceptual information shown to the end user. Instead of showing all the dimensions together, users can restrict their searches according to context and consequently avoid overinformation.

On the other hand, we have shown how constraints can be defined for KPs to facilitate knowledge extraction. This makes the terminographer's job easier and provides the end user with more explicitly related information. Needless to say, there is still much work to be done before achieving an automatic knowledge extraction system. For example, even if all KPs were tightly constrained, we are still analysing terms, and synonymic values may seem different concepts. That could be solved if knowledge representation and extraction processes were complementary. In this way, ontological systems could inform the extraction system and semantic annotation could be based on an already defined hierarchy. In any case, manual validation would still be necessary. Once we extract all the information we need to avoid noise and disambiguate patterns and dimensions, we are planning to manually annotate a significant part of the corpus to use it as a gold standard. In this way, automatic procedures could next be implemented and slowly refined.

References

Cimiano, P., Staab, S. (2005) 'Learning Concept Hierarchies from Text with a Guided Agglomerative Clustering Algorithm', *Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods,* Bonn.

Hanks, P., Pustejovsky, J. (2005) 'A Pattern dictionary for natural language Processing', *Revue Française de linguistique appliquée,* 10: 2, 63-82.

León Araúz, P. (2009) 'Representación multidimensional del conocimiento especializado: el uso de marcos desde la macroestructura a la microestructura', PhD dissertation, Universidad de Granada.

León Araúz, P., Magaña Redondo, P.J., Faber, P. (2009) 'Building the SISE: an environmental ontology', *Proceedings of the Towards e-Environment Conference*, Prague.

Marshman, E., L'Homme, M.C. (2006) 'Disambiguation of lexical markers of cause and effect', *Proceedings of the 15th European Symposium on Language for Special Purposes, LSP 2005*, Berna: Peter Lang, 261-285.

Pustejovsky, J., P. Hanks, Rumshisky, A. (2004) 'Automated induction of sense in context', *Proceedings of the 20th international conference on Computational Linguistics*, Geneva, 924-931.

Rumshisky, A., Hanks, P., Havasi, C., Pustejovsky, J. (2006) 'Constructing a Corpus-based Ontology using Model Bias', *Proceedings of FLAIRS 2006*, Melbourne Beach, Florida, 327-332.

Yeh, W., Barsalou, L.W. (2006), 'The situated nature of concepts', *American Journal of Psychology*, 119, 349-384.