Evaluation of EcoLexicon Images

Pilar León-Araúz, Arianne Reimerink

Department of Translation and Interpreting, Universidad de Granada

Buensuceso 11, 18071 Granada (Spain)

pleon@ugr.es, arianne@ugr.es

Abstract

The multimodal knowledge base EcoLexicon includes images to enrich conceptual description and enhance knowledge acquisition. These images have been selected according to the conceptual propositions contained in the definitional templates of concepts. Although this ensures coherence and systematic selection, the images are related to each specific concept and are not annotated according to other possible conceptual propositions contained in the image. Our aim is to create a separate repository for images, annotate all knowledge contained in each one of them and then link them to all concept entries that contain one or more of these propositions in their definitional template. This would not only improve the internal coherence of EcoLexicon but it would also improve the reusability of the selected images and avoid duplicating workload. The first step in this process and the objective of the research here described is to evaluate the images already contained in EcoLexicon to see if they were adequately selected in the first place, how knowledge is conveyed through the morphological features of the image and if they can be reused for other concept entries. This analysis has provided preliminary data to further explore how concept type, conceptual relations, and propositions affect the relation between morphological features and image types chosen for visual knowledge representation.

Keywords: multimodality, knowledge representation, visual knowledge patterns, resource evaluation

1. Introduction

EcoLexicon (ecolexicon.ugr.es) is a multimodal, multilingual terminological knowledge base (TKB) on the environment in which concepts are interrelated, based on the information extracted from a specialized domain corpus created for EcoLexicon (Faber et al., 2014; León Araúz et al., 2009). However, the way concepts are related can also be reflected in the graphical images depicting these concepts. For this reason, a visual corpus was also compiled to enrich conceptual description in EcoLexicon.

Currently, images are stored in association with concept entries according to the semantic content described in their definition. The definitions in EcoLexicon are based on templates that define category membership and describe the basic conceptual propositions. In this way, definitions have a uniform structure that directly refers to and evokes the underlying conceptual structure of the domain. These templates can be considered a conceptual grammar that thus ensures a high degree of systematisation (Montero and García 2004; Faber et al. 2007; Faber 2012). For example, for the definition of EROSION, the template includes the four basic relations of all natural processes: *is_a*, *has_agent*, *affects* and *has_result*. It also has an additional relation because it is a complex procedural concept, which can be divided into a sequence of steps: *has_phase*. For the selection of images, the basic conceptual propositions in the definitional template are used to select images which contain the same information to reinforce knowledge acquisition (Faber et al., 2007).

The definitional template provides a systematic means to select images for each concept entry in the TKB. It has been applied by our researchers to compile the visual corpus for over ten years (Faber et al., 2007). Images are thus regarded as a whole and are only linked to the concept itself. So far, we have shown how the same concept can be (and most often should be) represented through different images, depending on perspective, or the semantic content highlighted (Faber et al., 2007; Reimerink et al., 2010). However, the same image can also work for the representation of other related concept entries (e.g. an entity and the process through which it was formed, a concept and its parts, etc.). Images should thus be further dissected according to the features they possess (i.e. image type and other morphological characteristics) and the knowledge they convey (i.e. semantic content). For example, many images show several concepts in a specific background where they establish different relations that can be explicitly labeled or inferred from previous knowledge. In this sense, we believe that images should not be stored in the TKB as the representation of a concept, but as the representation of a set of conceptual propositions. They should be annotated according to semantic and morphological information and stored in a separate repository. Since each image activates several propositions and each proposition can be activated by different concepts, one image would then be linked to several concept entries. This would enhance the reusability of images, improving the consistency of the TKB and avoiding duplicating workload.

The aim of the research here described is to take the first step in this process of improving image selection and annotation: evaluate the images already contained in EcoLexicon to see i) if they were adequately selected in the first place; ii) how conceptual knowledge is conveyed through the morphological features of the image and iii) if they can be reused for other concept entries. The results of the evaluation will shed light on the relationship between the conceptual relations conveyed and the morphological features, or visual knowledge patterns (VKPs) used in images. In the future, the data obtained will help us to create the necessary annotation criteria for the existing and newly selected images in EcoLexicon.

2. Methodology

The images were analyzed in terms of (1) their adequateness in representing one or more conceptual propositions linked to the concept entry in which there are currently included, (2) the reusability of the image for other concept entries, (3) the semantic relations expressed in the image, (4) the concept types involved, and (5) the morphological features or visual knowledge patterns (VKPs) such as colour coding, referential background, arrows, labels, etc. used to convey the information. Two of our researchers studied all entries en EcoLexicon and then discussed the fulfilment of criteria and image description. An image was considered reusable if it contained any other concept or conceptual proposition that is or could be included in the TKB.

The images and the concepts they are linked to were exported to a spread sheet, where they were manually assigned a number according to the level of adequateness: 0, not at all; 1, partially; and 2, completely. Then another number was assigned according to the reusability of the image (0, no; 1, yes). Other columns included information on the image type (photograph, drawing – including maps and diagrams – or flow chart), semantic content (concepts and relations) and VKPs (labels, arrows, colours and their specific use). The assessment outcome is explained with the example in Figure 1 and Table 1.

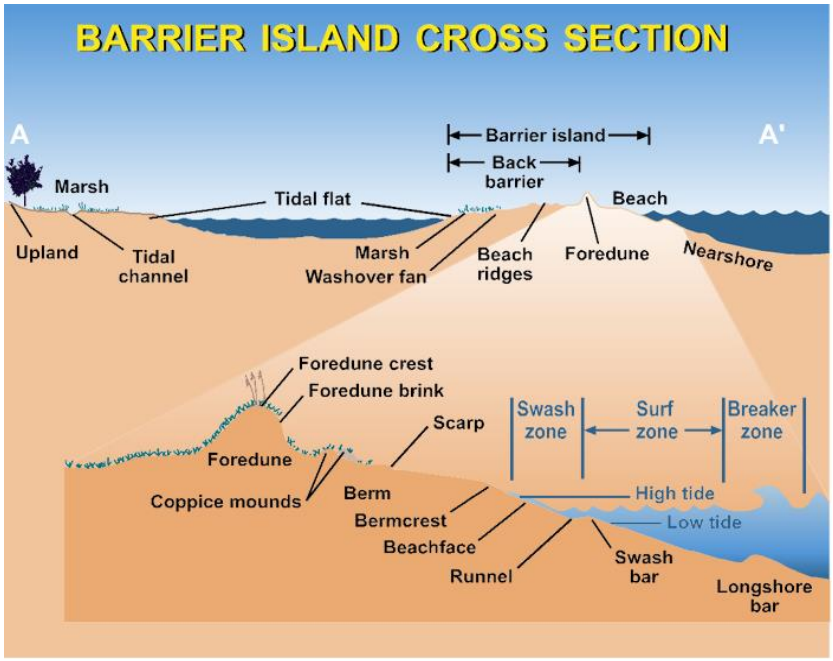

Figure 1: Drawing in concept entry BARRIER ISLAND

Concept	Image type	Conceptual propositions	VKPs	Adequate	Reusability
Barrier island	Drawing	Barrier island *has_location* nearshore Barrier island *has_part* back barrier/ foredune/ beach Barrier island *delimited_by* tidal flat/nearshore etc.	Static image Colours (referential similarity and distinction) Arrows (denomination of parts)	2	1 (foredune, tidal flat, surfzone, etc.)

Table 1: Assessment outcome for drawing in concept entry BARRIER ISLAND

The image in Figure 1 is found in the entry for BARRIER ISLAND and it represents several conceptual propositions. It has therefore been assigned a 2 (see Table 1), for it is fully adequate. The combination of image type (drawing) and VKPs (colours that provide a high level of referential similarity and arrows with labels) makes it especially adequate for the representation of the *part_of* relation. In this specific case, the image represents more conceptual propositions (*has_location* and *delimited_by*) because of the larger context in which the concept is shown. It can also be easily reused in entries of the parts of the concept and, again because of the larger context, in geographically related concepts such as TIDAL FLAT and SURF ZONE.

3. Results and Discussion

The results of our evaluation will be described and discussed in two subsections. The first one will show the data related to the adequateness of the images for the concept entry in which they are included and the possibility to use them in other concept entries. In this section examples of EcoLexicon will be included to

underline our discussion of the results. The second subsection explains how the different types of image, relation, concept and VKP are combined to convey knowledge in EcoLexicon images. Correlation graphs are used to visualize the data.

3.1 Adequateness and Reusability

Currently, EcoLexicon includes 3599 concepts and 1113 concepts have one or more images linked to their entry (31%). The total number of images amounts to 1698 of which 90.8% are adequate, 8.0% are partially adequate and 1.2% are not. Some are only partially adequate because the images are too small or unclear in the sense that you cannot see them properly, which is probably due to changes in format when introduced in EcoLexicon from other sources. The cases of complete inadequateness, normally due to misinterpretation of the conceptual information contained, must be discarded and new adequate images selected.

Although most images are adequate in the sense that they represent at least one conceptual proposition, sometimes several very similar images that do not add distinguishable conceptual knowledge are selected for the same entry in EcoLexicon. AQUIFER is a good example with four drawings with exactly the same information. Of these, the best one should be selected and linked to all the related concepts. The rest should be discarded. On the other hand, in the concepts CONFINED and UNCONFINED AQUIFER one image (Figure 2) that perfectly distinguishes these closely related concepts has been used.

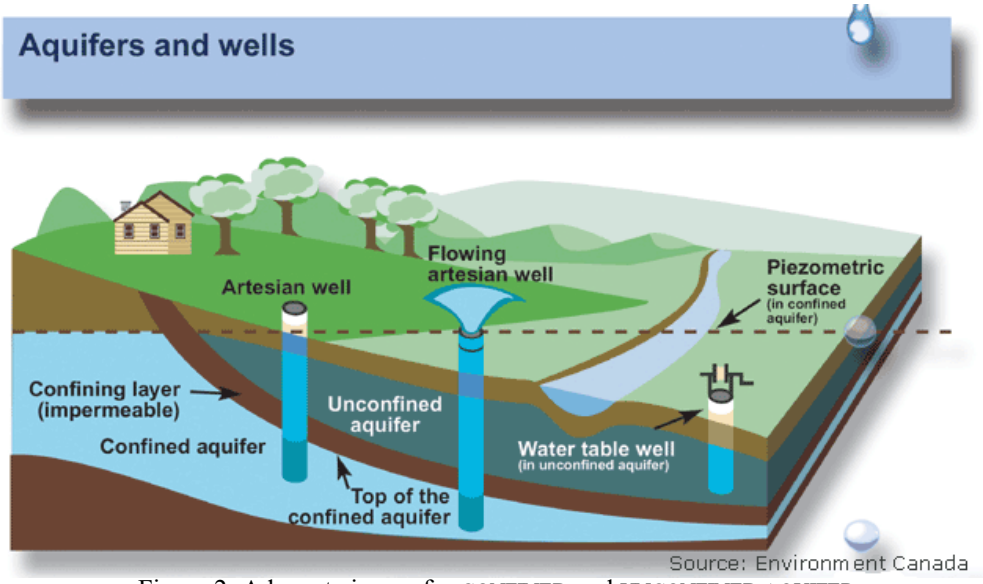

Figure 2: Adequate image for CONFINED and UNCONFINED AQUIFER

Most images are colour photographs representing *type_of* and *has_location* relations. These images cannot be reused often, because they represent the real world object. Some of them could be reused if we take into account that these real world objects are the result of certain processes, for example the concept BAR in Figure 3 is a colour photograph that represents the conceptual propositions BAR *has_location* RIVER MOUTH and BAR *result_of* LITTORAL SEDIMENTATION.

Through the annotation of all conceptual propositions in these images, they can be linked to the concept entries of the related processes – and not only to those of the entities resulting from them –, which would be a way to enhance reusability.

Instruments are often depicted by colour photographs as well. This is a problem because the most important conceptual relation for instruments is *has_function*, which cannot be easily represented by a photograph. A good example is DREDGING in Figure 4, where the *takes_place_in* and *has_instrument* relations are quite clear, but the *has_function* relation is not.

Figure 3: Adequate image for BAR *has_location* RIVER MOUTH and BAR *result_of* LITTORAL SEDIMENTATION

Figure 4: Adequate image for DREDGING *takes_place_in* SEA and *has_instrument* DREDGE

These images should therefore be combined with images that represent the process in which they participate, so that the *has_function* relation is made explicit. Another possibility to make the *has_function* relation explicit is to combine instruments with the output they provide, for example the concept entry METEOROGRAPH includes a photograph of the instrument and the concept METEOROGRAM includes a photograph of the output of the instrument. If we explicitly link both images to both concept entries, the *has_function* relation will be much clearer.

The number of photographs (51.7%) largely exceeds the number of drawings (28.8%) and flow charts (19.4%). There are several reasons to explain this.

Firstly, there are more objects than events in the knowledge base. Secondly, when the implementation of EcoLexicon began, quick deployment was considered more important than a coherent view towards image selection. Flow charts and drawings are more reusable than photographs and, if all conceptual propositions in the images are carefully annotated, they can be shown according to the specific perspective of the end-user providing more pertinent and coherent information. The entry for REEF is an example of how this can be done. The three drawings in Figure 5, for example, explain the phases of reef formation as well as the name of the subtype of reef in each phase. The image can be reused for each of the subtypes and in combination with a colour photograph of the real world entities, will provide the end-user with all the necessary information for comprehension: how each subtype is formed, how it relates to the other subtypes, and what they look like in reality.

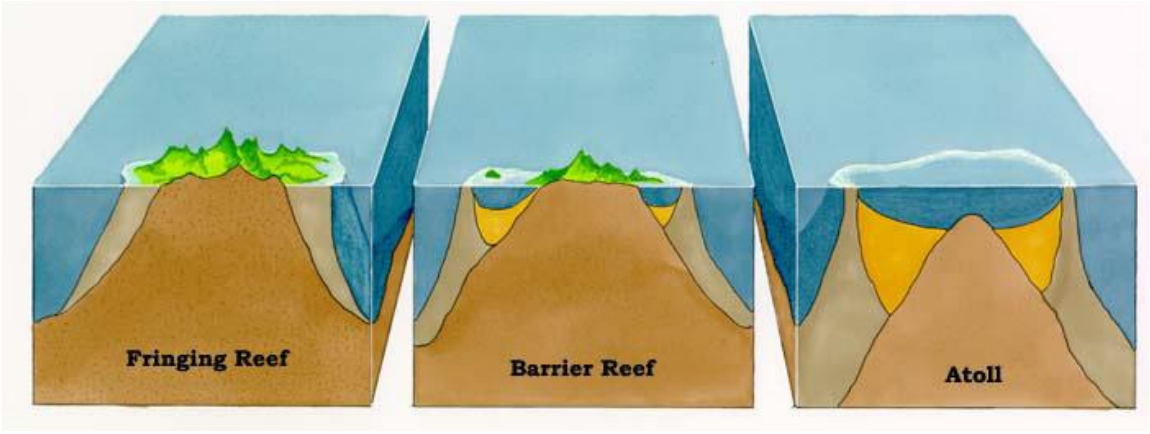

Figure 5: Adequate drawing for REEF *has_phase* FRINGING REEF/BARRIER REEF/ATOLL and FRINGING REEF/BARRIER REEF/ATOLL *type_of* REEF

Flow charts provide much conceptual information and can therefore be used to add new conceptual entries to EcoLexicon. However, this has not been done often.

Figure 6: Adequate image for conceptual propositions combining UPWELLING, COASTLINE, WIND and SURFACE WATER

For example, Figure 6 is only found in the entry for UPWELLING, but could have been reused in concept entries such as SURFACE WATER, WIND and CONTINENTAL SHELF. That way, an end-user would be able to understand the interaction among these entities no matter with which one the search started.

With the results of our study and its description of the images, all flow charts can be revised to search for new conceptual information. Moreover, annotating all conceptual propositions in the flow charts in detail as well as their VKPs will boost reusability.

3.2 Correlation between Image Types, Semantic Relations and VKPs

According to the results in Figure 7, processes are represented mostly with flow charts, whereas photographs and drawings are used to describe entities. A large number

of entities (over 40%) is also represented by flow charts. This is due to the fact that processes can affect physical entities and the latter can cause processes, thus their interaction is better conveyed in flow charts. Processes can also be depicted in a combination of drawings were different phases are shown in each one of them (e.g. Figure 5) and in photographs when the focus is on the result of the process (e.g. Figure 3). Properties (e.g. SOIL PERMEABILITY or ISOTROPIC) appear more often in drawings because you need labels to explicitly convey them, and labels are the most prototypical VKP in drawings (see Figure 10).

Not surprisingly, as flow charts are clearly preferred for representing processes, they are also the image type mostly used to convey procedural relations such as *result_of* and *causes* (Figure 8; only the most representative relations are shown). In turn, photographs are clearly more adequate for *type_of* and *has_location*, whereas drawings are used evenly for *type_of* and *has_location*, as well as *part_of* and *delimited_by*, all typical relations for the description of physical entities. Actually, *part_of* and *delimited_by* are specific to drawings.

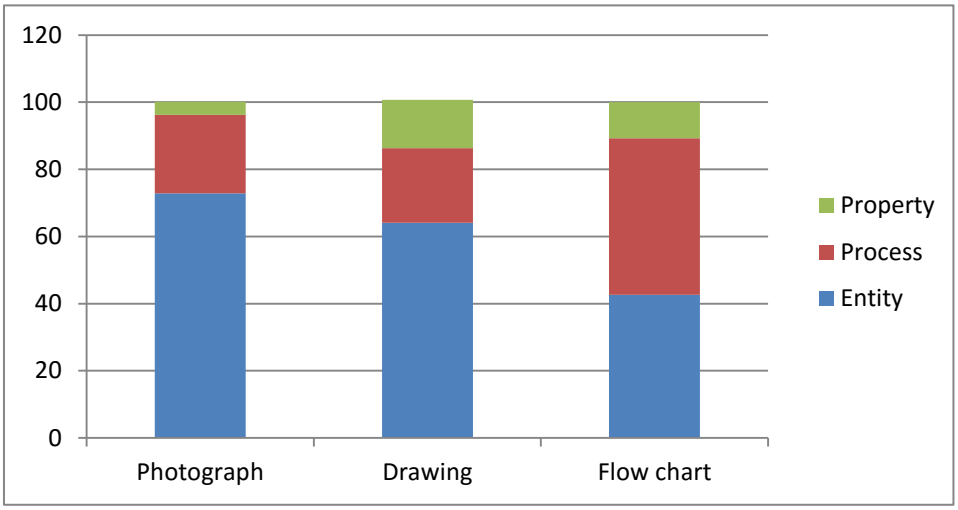

Figure 7: Correlation between image types and concept types

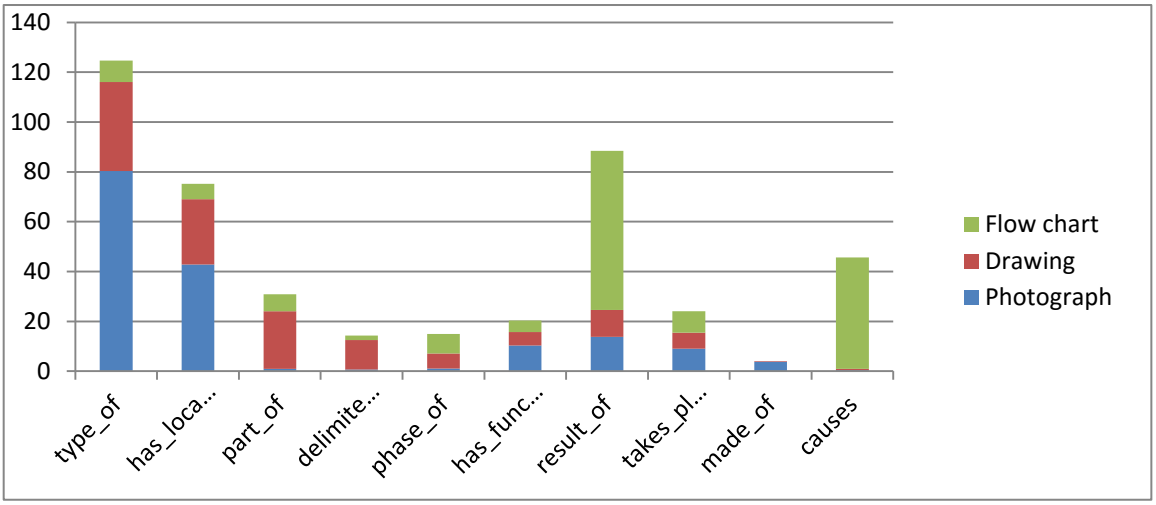

Figure 8: Correlation between conceptual relations and image types

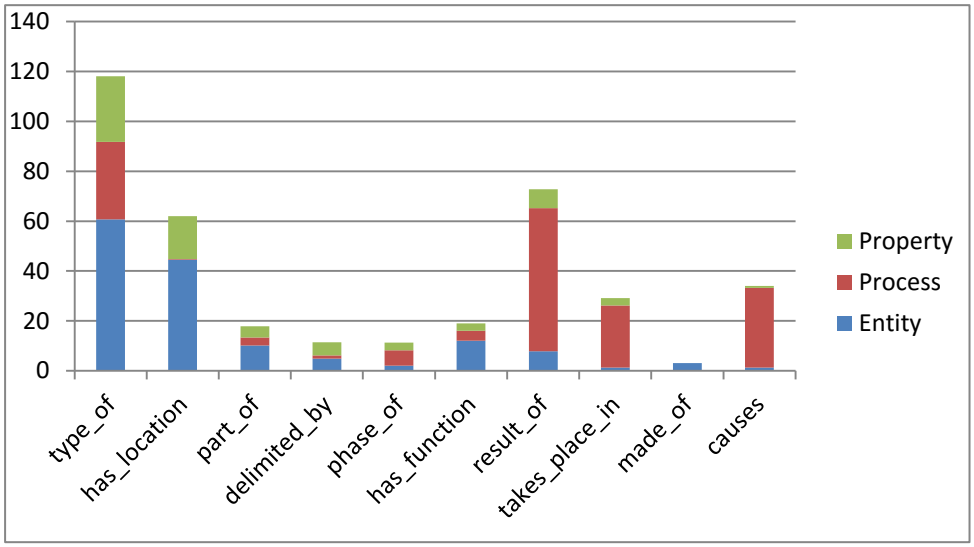

Figure 9: Correlation between conceptual relations and concept types

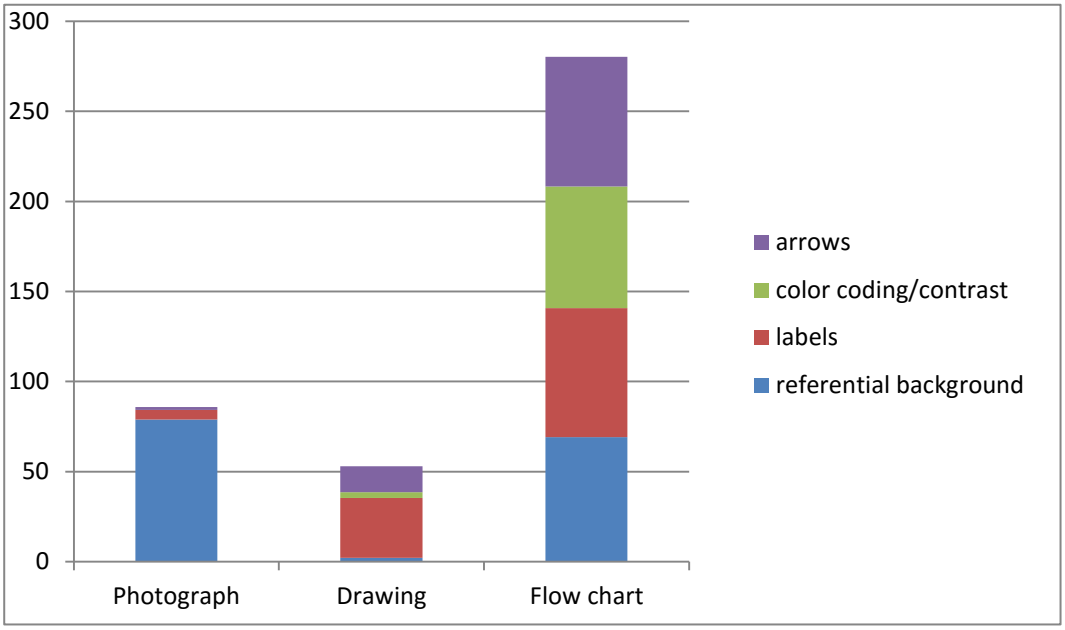

Figure 10: Correlation between image types and VKPs

Processes are mostly represented in images where relations such as *result_of*, *takes_place_in* and *causes* are present (see Figure 9). Entities, however, are found in images in combination with *type_of*, *has_location* and, to a lesser extent, *part_of* and *has_function*. *Has_function* and *part_of* are less represented because not all entities are functional nor do they have differentiated parts. Moreover, we have found that the *has_function* relation is not easy to convey in one image (see Figure 4). Properties are mostly present in images that convey *type_of* and *has_location*. They are also present, approximately in the same percentage, in combination with the other relations, except for *made_of*, which may be due to the surprising lack of data on this relation.

In Figure 10 only the most representative VKPs have been included. All of them are equally present in flow charts in EcoLexicon, they even appear all together in many of them. Photographs have clear referential backgrounds in over 75% of all cases and occasionally include labels. As previously mentioned, drawings are mostly characterized by labels and then arrows. The obvious conclusion we can draw from these results is that flow charts are the most interesting image type for the study of VKPs.

4. Conclusion

Much has been written regarding the importance of combining visual and textual information to enhance knowledge acquisition (Paivio, 1971, 1986; Mayer & Anderson, 1992). However, the combination of images and text still needs further analysis (Faber, 2012; Prieto Velasco, 2008; Prieto Velasco & Faber, 2012). An in-depth analysis of the features of images provides the means to develop selection criteria for specific representation purposes. The combination of conceptual content and image type based on morphological characteristics can be used to enhance the selection and annotation of images that explicitly focus on the conceptual propositions that best define concepts in a knowledge base. The analysis of EcoLexicon images has

provided the preliminary data to further explore how concept type, conceptual relations, and propositions affect the relation between VKPs and image types chosen for visual knowledge representation. Depending on how the annotation process evolves, new data will provide the clues for future content-based image description and retrieval.

5. Acknowledgements

This research was carried out as part of project FF2014-52740-P, *Cognitive and Neurological Bases for Terminology-enhanced Translation* (CONTENT), funded by the Spanish Ministry of Economy and Competitiveness.

6. Bibliographical References

Faber, P. (Ed.) (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/ Boston: De Gruyter Mouton.

Faber, P., León-Araúz, P. and Reimerink, A. (2014). Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era, Educational Linguistics* 19, ed. E. Bárcena, T. Read and J. Arhus. Berlin, Heidelberg: Springer.

Faber, P., León Araúz, P., Prieto Velasco, J. A., and Reimerink, A. (2007). Linking images and words: the description of specialized concepts. *International Journal of Lexicography* 20(1), pp. 39–65, doi:10.1093/ijl/ecl038.

León Araúz, P., Reimerink, A., and Faber, P. (2009). Knowledge extraction on multidimensional concepts: corpus pattern analysis (CPA) and concordances. In *8ème Conférence Internationale Terminologie et Intelligence Artificielle*. Toulouse.

Mayer, R. E. and Anderson, R. B. (1992). The instructive animation: helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology* 84(4), pp. 715–726.

Montero-Martínez, S. and García de Quesada, M. (2004). Designing a corpus-based grammar for pragmatic terminographic definitions. *Journal of Pragmatics* 36(2), pp. 265–291.

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental Representations: A Dual-Coding Approach*. New York: Oxford University Press.

Prieto Velasco, J. A. (2008). *Información Gráfica y Grados de Especialidad en el Discurso Científico-Técnico: Un Estudio de Corpus*. PhD Thesis, University of Granada.

Prieto Velasco, J. A. and Faber, P. (2012). Graphical Information. In P. Faber (Ed.) *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin/Boston: De Gruyter Mouton, pp. 225–248.

Reimerink, A., García de Quesada, M. and Montero Martínez, S. (2010) Contextual information in terminological knowledge bases: A multimodal approach. *Journal of pragmatics* 42(7), pp. 1928-1950.

6.1 Sources of Images

Figure 1:
http://www.beg.utexas.edu/UTopia/contentpg_images/gloss_barrier_island2.jpg

Figure 2:
http://water.usgs.gov/gotita/earthgwaquifer.html

Figure 3:
http://www.mappinginteractivo.com/plantilla-ante.asp?id_articulo=464

Figure 4: http://www.dragadoshidraulicos.com/

Figure 5:
http://www.biosbcc.net/ocean/marinesci/04benthon/crform.htm

Figure 6:
http://cordellbank.noaa.gov/images/environment/upwelling_470.jpg