Predicate-Argument Analysis to Build a Phraseology Module and to Increase Conceptual Relation Expressiveness

Arianne Reimerink(✉) and Pilar León-Araúz

Department of Translation and Interpreting, University of Granada, Granada, Spain
{arianne,pleon}@ugr.es

Abstract. EcoLexicon, a multilingual and multimodal terminological knowledge base (TKB) on the environment, needs improvements: more expressive non-hierarchical relations and a phraseology module consistent with knowledge representation in the other modules of the TKB. Both issues must be addressed by analyzing predicate-argument structure in text. In this paper, we explain our methodology for predicate-argument analysis with the case study on the conceptual relation *affects*. We take a semi-automatic approach to extract term-verb-term collocates with Sketch Engine [1]. Then the verbs are classified according to the lexical domains proposed by Faber & Mairal [2] and the arguments in conceptual categories based on the knowledge contained in EcoLexicon. To validate the lexical domains and conceptual categories, an automatic clustering method based on word2vec [3] is applied. The analysis of verbs and arguments contributes to the refinement of our semantic relations and categories as well as to the population of the phraseological module.

Keywords: Predicate-argument analysis · Phraseology · Conceptual relation expressiveness

1 Introduction

EcoLexicon[1] is a multilingual and multimodal terminological knowledge base (TKB) on the environment. In the construction of the TKB, two different but related problems have arisen. On the one hand, we are working on the design of a phraseology module that is consistent with the knowledge extraction and representation methodology based on triplets, or conceptual propositions, in EcoLexicon [4]. On the other hand, the semantic expressivity of some of the conceptual relations in the TKB's semantic networks should be improved. For instance, conceptual propositions such as EROSION *affects* LANDFORM would be more meaningful if the relation was *reduces* instead of *affects*. However, the phraseological module of the TKB should also contain other verbs lexicalizing and specifying the nuclear meaning of reduction (e.g. *carve*, *degrade*, *erode*, etc.) as well as other terms that can also fill the

[1] ecolexicon.ugr.es.

R. Mitkov (Ed.): Europhras 2017, LNAI 10596, pp. 176–190, 2017.
https://doi.org/10.1007/978-3-319-69805-2_13

slots of these arguments (e.g. *weathering*, *cliff*, etc.). To solve these problems, the first step is to analyze predicate-argument structure in real text.

We understand phraseology from a broad perspective as all word combinations with certain stability [5, 6]. According to Rundell [7] (vii), collocations are as important as grammar since they make speakers/writers sound fluent. In specialized domains, they are perceived by language users to contribute to the domain-specific flavor of special languages [8]. In this line, recent studies have highlighted the importance of verbs, their collocations and argument structure in specialized terminology [9, 10], but there are currently few terminographic resources that incorporate them (exceptions are DiCoInfo and DiCoEnviro [11] and DicSci [12], for example). If terminological knowledge bases (TKBs) want to be truly helpful for specialized writing, phraseological information should be added in a consistent and user-friendly way.

In an attempt to connect the description of predicative units to the knowledge structure [11: 89] of EcoLexicon and make the phraseological module consistent with the conceptual module, it should be based on the same principles. Therefore, we propose a design based on the categorization of term-verb-term collocates reflecting the different lexicalizations of conceptual propositions. In this way, semantic relations can be further specified according to specialized predicates. In turn, phraseological templates can be generalized based on the semantic types related in conceptual networks. However, these semantic types still need to be extracted in a consistent way. In this paper, we explain our methodology for predicate-argument analysis with the case study on the conceptual relation *affects*. The analysis of verbs and arguments will contribute to the refinement of our conceptual relations and categories as well as to the population of the phraseology module.

In Sect. 2, the EcoLexicon TKB is described in more detail. In Sect. 3, the methodology for predicate-argument analysis is explained. Section 4 describes how the results of predicate-argument analysis affect the representation of conceptual networks and phraseology module design in EcoLexicon. In Sect. 5, word2vec clustering is used to validate the conceptual categories and lexical domains defined in Sect. 3 and to extract new seed terms for further analysis. Conclusions are drawn and future work is proposed in Sect. 6.

2 EcoLexicon

EcoLexicon is a multilingual and multimodal terminological knowledge base (TKB) on the Environment. It currently contains 3,601 concepts and 20,211 terms in English, Spanish, French, German, Russian and Modern Greek. It is the practical application of Frame-based Terminology (FBT), a cognitively-oriented theory of specialized knowledge representation that applies certain features of Frame Semantics [13] to structure specialized domains and create non-language-specific representations. FBT focuses on: (i) conceptual organization; (ii) the multidimensional nature of specialized knowledge units; and (iii) the extraction of semantic and syntactic information through the use of multilingual corpora. FBT operates on the premise that specialized knowledge units

activate domain-specific semantic frames that are in consonance with users' background knowledge [14].

EcoLexicon is an internally coherent information system, which is organized according to conceptual and linguistic premises at the macro- as well as the micro-structural level. It targets users such as translators, technical writers, and environmental experts who need to understand specialized environmental concepts with a view to writing and/or translating domain specific texts. Users interact with EcoLexicon through a visual interface (see Fig. 1). The top horizontal bar gives users access to the term/concept search engine. The vertical bar on the left of the screen provides information regarding the search concept, namely its definition, term designations, associated resources, general conceptual role, and phraseology. The center area has tabs that access the following: (i) the history of concepts/terms visited; (ii) the results of the most recent query; (iii) all the terms alphabetically arranged; (iv) the shortest path between two concepts; and (v) concordances for a term. On the center of the screen, the conceptual map is shown as well as the icons that allow users to configure and personalize it for their needs. The standard representation mode shows a multi-level semantic network whose concepts are all linked in some way to the search concept, which is at its center [15].

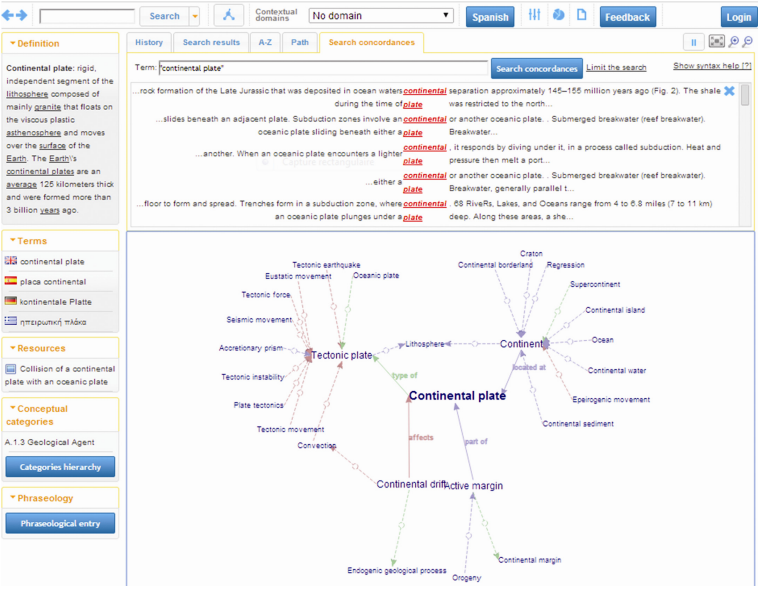

Fig. 1. EcoLexicon user interface.

2.1 Conceptual Relations

Conceptual description in EcoLexicon is based on concept types and their relational behavior. We use a fixed set of conceptual relations (see Table 1) that have been defined according to coherent and systematic criteria in order to make EcoLexicon a consistent resource at its different representational levels.

Table 1. EcoLexicon inventory of conceptual relations.

Relation category	Relation	Example
Generic-specific	Type_of	GROYNE *type_of* COASTAL DEFENSE STRUCTURE
Part-whole	Part_of	SPILLWAY *part_of* DAM
	Made_of	GROYNE *made_of* WOOD
	Delimited_by	STRATOSPHERE *delimited_by* STRATOPAUSE
	Located_at	GROYNE *located_at* COAST
	Takes_place_in	LITTORAL DRIFT *takes_place_in* SEA
	Phase_of	PUMPING *phase_of* DREDGING
Non-hierarchical	Affects	GROYNE *affects* LITTORAL DRIFT
	Causes	WATER *causes* EROSION
	Result_of	ACCRETION *result_of* SEDIMENTATION
	Has_function	AQUIFER *has_function* WATER SUPPLY
	Studies	POTAMOLOGY *studies* SURFACE CURRENT
	Measures	PLUVIOMETER *measures* PRECIPITATION
	Effected_by	DREDGING *effected_by* DREDGER

For instance, meronymy has been split up into six different relations, since not all parts interact in the same way with their wholes. In the same way the expressiveness of meronymy has been increased by splitting it up into six different relations, some non-hierarchical relations need more specification as well. This is the case of *affects*. This conceptual relation has become a catch-all relation in EcoLexicon, where for example ABLATION *affects* GLACIER and WATER DENSITY *affects* WATER. In the first case, the relation expressed would be that GLACIER is a *patient* of ABLATION, whereas in the second WATER DENSITY is an *attribute* of WATER. An example of a predicate-based relation where more expressiveness would be needed is BREAKWATER *affects* BEACH. In this case, it would be interesting to know how one affects the other, as the consequences of having a breakwater may be positive for a beach (e.g. *protect*) in some cases and negative in others (e.g. *erode*). One case where the relation is not only inexpressive but also confusing is where EcoLexicon tells its users that ULTRAVIOLET RADIATION *affects* OZONE and OZONE *affects* ULTRAVIOLET RADIATION. We will discuss the latter in detail in Sect. 3. Refining the *affects* relation would greatly improve knowledge acquisition by non-expert end users.

2.2 Phraseology Module

The phraseology module of EcoLexicon is currently under construction. Until recently, phraseological information was stored at the term level, where verbs were related to arguments contained in EcoLexicon as shown in Fig. 2. The verbs that collocate with the term were classified according to the lexical domain and subdomain based on Faber & Mairal [2] (see Sect. 3). The partial phraseology entry for hurricane shown in Fig. 2 includes the lexical domains Action and Change. Within Action, the subdomain "to come against something with sudden force" includes the definition of the phraseological pattern and the verbs *hit*, *batter*, *strike*, and *blast*. By clicking on *hit*, the user accesses

four usage examples as well as a note section with information about meaning restrictions. In this case, the note states, among other things, that the natural force is usually an atmospheric agent, water agent, natural disaster or atmospheric condition [10].

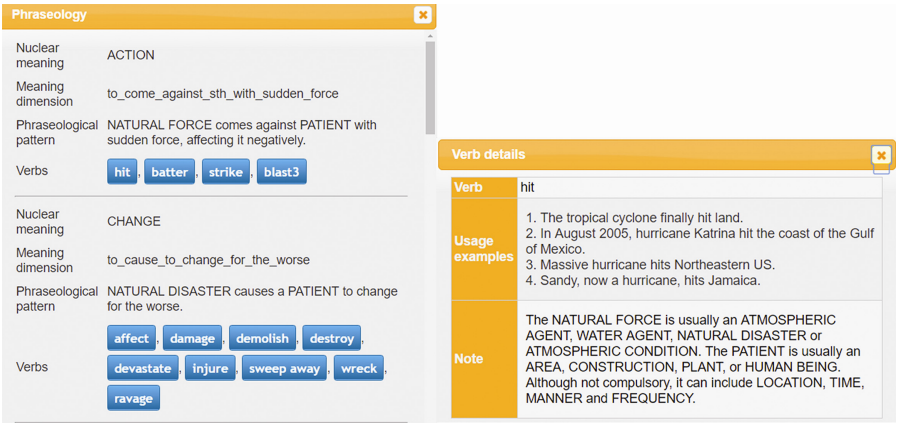

Fig. 2. Current phraseology entry for *hurricane*.

We intend to improve the phraseology module of EcoLexicon by creating phraseological templates that represent a generalization of predicate-argument structure, where lexical domains of verbs are related to the conceptual categories of their arguments.

3 Case Study: *Affects*

The case study on *affects* was carried out along several phases. Firstly, term-verb-term combinations that represent *affects* were analyzed in a 67-million-word specialized corpus on the environment. This was done by selecting all conceptual propositions linked through the *affects* relation in EcoLexicon in need of a predicate-based refinement, which means that patient and attribute-based propositions were ruled out for this study. Then, the verbs lexicalizing the extracted triplets in the corpus were classified into lexical domains. Thirdly, to increase the expressiveness of the relation *affects*, a verb was chosen within the higher level hierarchy of the lexical domain in order to relabel the relation. The rest of the verbs were collected to be stored in the phraseological module. The terms in the argument slots of the verbs were classified in conceptual categories and, finally, the predicate-argument structure of the verbs and their arguments was described for the phraseology module.

3.1 Corpus Analysis

We started out with the terms *ozone* and *ultraviolet radiation* to clarify the confusing example described in Sect. 2.1 where one affects the other and vice versa. We used

Sketch Engine [1] to analyze the corpus, more specifically wordsketch and the simple query with context restrictions (see Fig. 3).

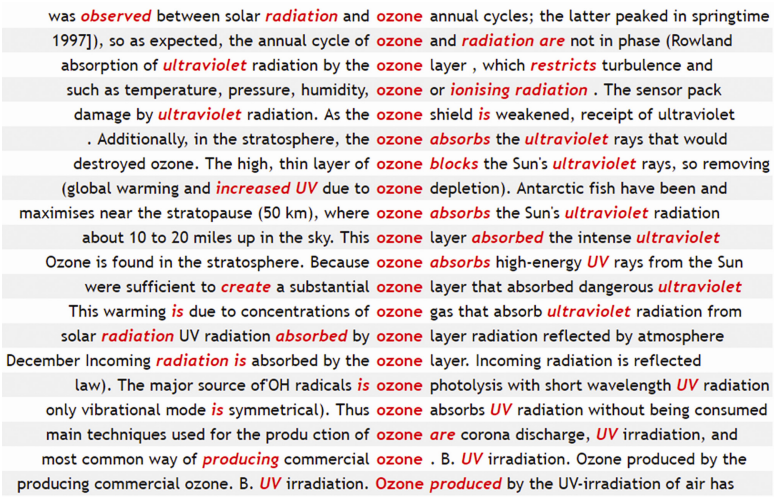

Fig. 3. Sample concordances of *ozone* + verb + *UV*, *ultraviolet*, *radiation*, *rays*.

In the concordances, combinations of *ozone* and *ultraviolet radiation* or *UV* or *rays* were found with verbs such as *absorb*, *filter out*, *block*, and *shield*. In this context, ozone was found as part of multiword terms such as *ozone layer*, *stratospheric ozone*, and *ozone shield*. For how *ultraviolet radiation* affects *ozone*, we found instances of ultraviolet radiation interacting with, *creating* or *destroying* ozone. Ultraviolet radiation can create as well as destroy ozone, because ozone is a very instable molecule.

Then we broadened our search to find combinations of *ozone* with different verbs with the basic underlying conceptual meaning of affect and the arguments that go with them. Ozone combines with verbs such as *shield* and *protect*, but also with *damage* and *irritate*. From the concordances with these verbs, it became clear that stratospheric zone (the ozone in the ozone layer) *shields* and, therefore, *protects*, whereas tropospheric (or ground-level) ozone *damages* and *irritates* (see Fig. 4). The second argument for both cases can be *Earth*, the *Earth's surface*, *wildlife*, *us, environment*. However, only in the case of ground-level ozone damages or irritates, the second argument refers to human health (*health*, *respiratory system*, *eyes*, *nose*, etc.).

To identify conceptual categories for the arguments and define lexical domains, the corpus queries were further broadened. For example, the verb *absorb* was queried in combination with *radiation*, *energy*, *rays*, *sunlight*, etc. The arguments found were all atmospheric components (*ozone*, *water vapor*, *carbon dioxide* and *greenhouse gases*; see Fig. 5). In another query, we combined synonyms of affect (*influence*, *damage*, *change*, *affect*) with second arguments such as *Earth*, *climate*, *environment* and *health*. In this case, we found greenhouse gases (*ozone*, *carbon dioxide*, *methane*, *nitrous oxide*, etc.) as a more specific semantic category.

warming, sea-level rise, and *reductions* in the **stratospheric ozone** protection *shields* . It was now imperative
key gases. 4.6 *Stratospheric* Chemistry. **Stratospheric ozone** *protects* life on the surface of the Earth
not as dramatically as in the *Antarctic* . **Stratospheric ozone** *protects* life on the surface of the Earth
molecular oxygen and atomic *oxygen* . There, **stratospheric ozone** provides a protective *shield* against the
appear to be involved with *ozone* depletion. **Stratospheric ozone** *shields* the earth from solar ultraviolet
health *problems* and *damage* forests and crops. **Ground-level ozone** affects the respiratory system, aggravating
chief *contributors* to ozone production. **Ground-level ozone** adversely affects health and *damages* the
They form photochemical *oxidants* (including **ground-level ozone**) that affect health, *damage* materials,
harmful ultraviolet (*UV*) radiation. However, **ground-level ozone** can *irritate* the eyes, noses, throats,
environment. Toxic *air* pollutants, rain and **ground-level ozone** can *damage* trees, crops, wildlife, lakes

Fig. 4. Stratospheric ozone protects and shields, ground-level ozone damages and irritates.

greenhouse gases. Atmospheric gases that **absorb** and reflect long-wave *radiation* , causing
circulation of *air* . latent heat. *Energy* being **absorbed** from the air during changes from water
and through photochemical *reactions* , that **absorbs** ultraviolet *radiation* from the sun, an
greenhouse gases. Atmospheric gases that **absorb** and reflect long-wave *radiation* , causing
circulation of *air* . latent heat. *Energy* being **absorbed** from the air during changes from water
and through photochemical *reactions* , that **absorbs** ultraviolet *radiation* from the sun, an
but most *reaches* the earth, where it is **absorbed** and re-emitted as long-wave *energy* , also
of our *atmosphere* , the ozone layer, that **absorbs** the dangerous part of the sun's *radiation*
that takes *place* as the sun's *energy* is **absorbed** varies greatly. The reason for this comes
infrared *radiation* rising from the surface is **absorbed** by CO2 in the middle levels of the atmosphere
world, and the *data* he used for how gases **absorbed** *radiation* were far from reliable. Nevertheless
quantity of *gas* . The reason was that CO2 **absorbs** *radiation* only in specific bands of the
difference . Moreover, water vapor already **absorbed** infrared *radiation* in the same region of
took to studying how CO2 in the *atmosphere* **absorbed** infrared *radiation* , as an adjunct to his

Fig. 5. Sample concordances of *absorb* with *radiation*, *energy*, *rays*, *sunlight*, etc.

3.2 Verb Classification

To classify the verbs found during corpus analysis in broader semantic categories, we applied the lexical domains and subdomains defined by Faber & Mairal [2]. These authors propose a model for lexical classification based on the distinction between syntagmatic and paradigmatic relations, where the most prototypical verbs are those that have the largest combinatory potential from a semantic point of view. They applied the model to over 10,000 verbs of the English language and the lexical domains they propose are: Existence, Movement, Position, Contact, Change, Perception, Cognition, Possession, Action, Feeling, Speech, Sound, and Light. Below, part of the Possession lexical domain hierarchy is reproduced with some example verbs in square brackets [2: 291].

We classified the verbs *absorb* (as in ozone absorbs ultraviolet radiation), *filter out*, and *block* within the lexical domain Possession and its subdomain "to come to have something". We considered *shield* and *protect* meronymic extensions of *block*. The verb *create* (as in ultraviolet radiation creates ozone) was classified in Existence in the subdomain "to cause something to exist". *Destroy* was included in Existence as well of course, but in the subdomain "to cause something to stop existing". *Damage*, *irritate*, and *harm* were classified in Change under "to cause something to change making it worse".

12 Possession
- 12.1 To have something [*possess*, *own*, *hold*]
  - 12.1.1 To come to have something [*get*, *obtain*]
    - 12.1.1.1 To get something as a result of force/skill [*take*, *capture*]
    - 12.1.1.2 To get something through effort/as a reward [*gain*, *earn*]
    - 12.1.1.3 To get something after it has been given/sent to you [*receive*]
    - 12.1.1.4 To get a large number of things over a period of time [*collect*, *accumulate*]
    - 12.1.1.5 To get something back after it has been lost/stolen [*recover*]
  - 12.1.2 To continue to have something [*keep*, *save*]
    - 12.1.2.1 To have something within as a part [*contain*, *include*]
    - 12.1.2.2 To cause something to have something as a part [*include*, *incorporate*]
      - 12.1.2.2.1 To not include [*omit*, *exclude*]
  - 12.1.3 To stop having [*lose*]

…

To increase the expressiveness of the conceptual relation *affects*, we then chose a verb from the same lexical subdomain that we felt expressed the relation more specifically but would still be applicable in other cases. For Possession ("to come to have something"), we chose *obtain*, for Existence ("to cause something to exist"), *create*, for Existence ("to cause something to stop existing"), *destroy*, and for Change ("to cause something to change making it worse"), *damage*. Further research will show if these choices are the most adequate for the environmental field.

3.3 Argument Classification

From the corpus analysis, several conceptual categories were deducted for the arguments in the predicate-argument structures found. For example, for the lexical domain Possession ("to come to have something"), the Agent was an atmospheric component with members such as *nitrogen*, *oxygen*, *argon* and the so-called trace gases (*stratospheric ozone*, *carbon dioxide*, *methane*, and *nitrous oxide*). The Patient in combination with this same lexical domain was some type of energy: *ultraviolet radiation*, *sunlight*, *sunrays*, *photons*, etc. For Existence and Change, we found the same Agents and Patients. The Agents were ozone-depleting substances, such as greenhouse gases (*water vapor*, *carbon dioxide*, *methane*, *nitrous oxide*, *chlorofluorocarbons*) and the Patients were stratospheric ozone, with its term variants *ozone layer*, *ozone*, *ozone shield*, etc., and the environment, including *Earth*, *climate* and *living beings*. The results are summarized in Table 2.

Table 2. Summary of argument and verb classification.

Argument 1 [Agent]	**Verb**	**Argument 2 [Patient]**
Conceptual category	Lexical domain	Conceptual category
Atmospheric component nitrogen oxygen argon trace gas: stratospheric ozone carbon dioxide methane nitrous oxide	Possession obtain retain absorb filter out block	Energy ultraviolet radiation sunlight sunrays photons
Ozone-depleting substance greenhouse gas: water vapor carbon dioxide methane nitrous oxide chlorofluorocarbons	Existence[a] destroy Change deplete degrade damage	Atmospheric component stratospheric ozone Environment Earth climate living beings humans health animals plants
Energy ultraviolet radiation sunlight sunrays photons	Existence[a] destroy Existence[b] Create Existence[a]	Atmospheric component stratospheric ozone Ozone depleting substance greenhouse gas water vapor carbon dioxide methane nitrous oxide chlorofluorocarbons

[a]Subdomain "to cause something to stop existing"
[b]Subdomain "to cause something to exist"

4 Applying Results

4.1 Conceptual Network Modification

The increased expressiveness of the conceptual relation *affects* will be applied to the conceptual networks of EcoLexicon. Currently, the network of ultraviolet radiation, for instance, shows ULTRAVIOLET RADIATION *affects* OZONE-DEPLETING SUBSTANCE and ULTRAVIOLET RADIATION *affects* STRATOSPHERIC OZONE (see Fig. 6). These relations will be more expressive when changing them to the verbs chosen in Sect. 3.2 (see Fig. 7).

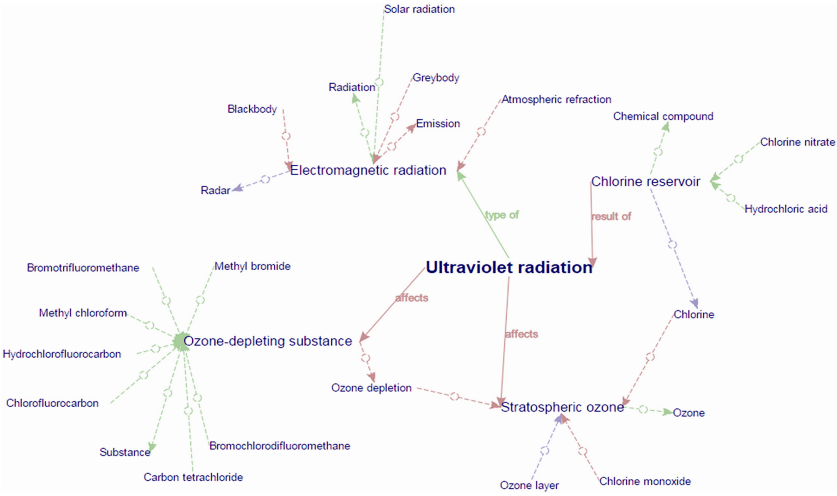

Fig. 6. ULTRAVIOLET RADIATION with inexpressive *affects* relation.

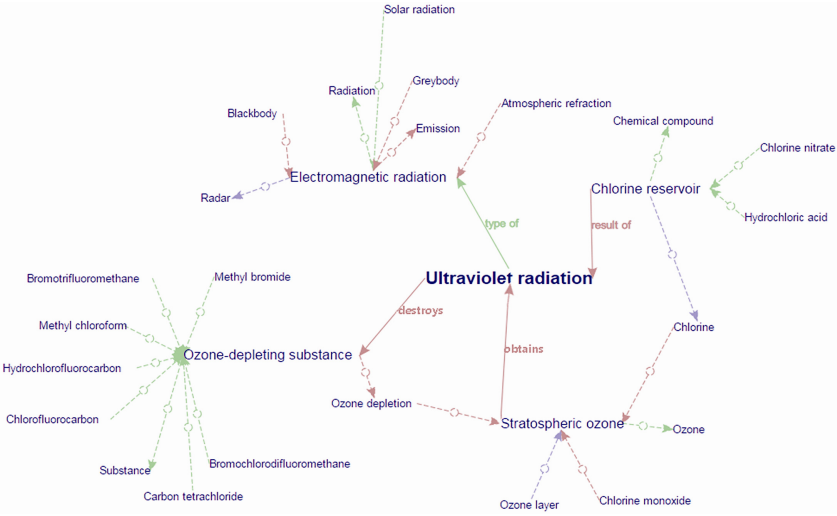

Fig. 7. ULTRAVIOLET RADIATION with increased expressiveness: *destroys* and *obtains*.

4.2 Phraseology Module Workflow

For detailed predicate-argument description in EcoLexicon's future phraseology module, the following workflow has been developed (see Fig. 8). First a term is selected within a concept entry, for example *tropospheric ozone* in the concept entry TROPOSPHERIC OZONE. Then a conceptual category is chosen for that concept (Step 2 in Fig. 8). After that, the lexical domain of the verb is selected (for example Possession) and its subdomain ("to come to have something"). In step 4, the role of the argument of the term in

step 1 is defined. It is not until step 5 that the actual verb is introduced (*absorb* in this case). In steps 6, 7 and 8, the same is done for the second argument. In step 9, example sentences are chosen from the corpus that express the predicate-argument structure. Depending on each specific case, steps can be repeated when necessary.

1. Select term1 within concept: tropospheric ozone
2. Conceptual category: greenhouse gas → gas → chemical → substance
3. Lexical domain verb: Possession (to come to have something)
4. Select semantic role term1: Agent
5. Add verb: absorb
6. Select conceptual category argument2: radiation → energy
7. Select semantic role argument2: Patient
8. Select or introduce argument2: ultraviolet radiation
9. Example sentences
10. Repeat steps when necessary

Fig. 8. EcoLexicon's phraseology module workflow.

In this way, the workflow forces consistency as well as a perspective towards phraseology that goes beyond the term level. Verbs are explicitly linked to their lexical domains, and arguments are linked to conceptual categories, creating a phraseological template. Apart from consistency and structural generalization, this workflow avoids duplicating work, as at each step, if the phraseological structure is the same for the argument-verb-argument combination, the conceptual categories of the arguments and the lexical domain of the verb will provide lists of all the items included. So, if we chose the lexical domain Possession ("to come to have something"), we can introduce *absorb* and *filter out* at the same time, as the arguments for both verbs coincide.

5 Semantic Categories and Lexical Domains Validation: Word2vec

In this section, we explore the possibilities of applying a word-space model to validate the semantic categories and lexical domains extracted in Sect. 3. The word-space model is a spatial representation of word meaning. What makes the word-space model unique in comparison with other geometrical models of meaning is that the similarities between words are automatically extracted from language data by looking at empirical evidence of real language use. Words with similar distributional properties are put in similar regions of the word space, so that proximity reflects distributional similarity [16: 21]. The idea behind this is the so-called distributional hypothesis: words with similar distributional properties have similar meanings [idem]. We used the word-space model word2vec [3] provided by Python's open source vector space topic modeling toolkit Gensim[2]. We chose the continuous bag-of-words (CBOW) hierarchy and applied it to our 67-million-word specialized corpus on the environment. Preprocessing included underscoring multiword terms contained in EcoLexicon present in the corpus and tokenizing with NLTK Toolkit[3].

[2] https://radimrehurek.com/gensim/models/word2vec.html.
[3] http://www.nltk.org/.

We created the model and then used the most similar method to obtain clusters of those vectors that are most similar to the search term with varying numbers of 20 or 40 (topn = 20 or topn = 40) different most similar vectors. The results are shown in Table 3. Searches were based, on the one hand, on the arguments *ozone* and *ultraviolet radiation* as well as those that were found accompanying any of the latter with similar predicates. In this way, the phraseological template of each of these terms could be rapidly enriched by adding new arguments belonging to the same conceptual category (note that most terms shown in the clusters are synonyms, hyponyms or cohyponyms of the search term). On the other hand, searchers were also based on the predicates found between ozone and ultraviolet radiation in order to enrich the lexical domains.

Search 1 and 2 provide clusters that coincide quite well with the items we included in the conceptual categories in Sect. 3.3. Search 1, for example, obtains all term variants for *ozone*, the term *greenhouse gases* as well as all of its hyponyms, and other atmospheric components. However, from the cluster only, it would be difficult to deduce the difference in meaning and behavior of stratospheric and tropospheric ozone. Search 3 provides a cluster with names of other planets and astronomical bodies, very different from the category of entities damaged by greenhouse gases we found in Sect. 3. Earth, climate and environment do not seem to be in the same category according the cluster. When looking for other entities damaged by greenhouse gases, *plant* (search 4) and *animal* (search 5), *tree* and *animal* are present in the former and *adult* and *human* in the latter, but the cluster still does not seem to fit our purposes. However, when using *human* as search word (search 6), the cluster includes tokens such as humankind, climate change, health, and global warming, which does give an accurate idea of how these can be included in the same category of the environment.

For the validation of the lexical domains of the verbs, the results of clustering seem to partially coincide with the results of Sect. 3.2, but the clusters would need refining to clearly show certain conceptual differences. For example in search 6 (*absorb*), *penetrate*, *enter*, *retain*, *consume*, *sequester* and *trap* all express the meaning of the lexical subdomain "to come to have something". According to the word2vec results, the opposite meaning is part of the same cluster as well, with verbs such as *emit*, *lose*, *dissipate*, and *radiate*, which are still part of the lexical domain Possession, but not of the same subdomain. Other verbs in the cluster are more closely related to the lexical domain Change, for example *attenuate* and *deplete*. Similar conclusions can be drawn from the other clusters. Further research will be necessary to refine the clustering technique, or to compare word2vec to other clustering algorithms such as Brown clustering [17] or Hierarchical Dirichlet Processing [18].

Apart from coinciding results that provide an approximate validation of the lexical domains and conceptual categories, clustering with word2vec also provides new members for each class, which can be added to the lexical domain or conceptual category in question and therefore enrich the module in a coherent and structured way. For example, *trap* and *sequester* can be added to the lexical domain Possession ("to come to have something"). These new members can then be used as seed words for further research. Thus, combining automatic procedures with manual queries to the corpus facilitates the population of the phraseological module and at the same time ensures coherence during the parallel development of the conceptual module.

Table 3. Clusters for search words in EcoLexicon corpus.

Search	Search word	Cluster
1	Ozone (topn = 20)	O3, stratospheric ozone, tropospheric ozone, CO2, atmospheric, aerosol, CH4, stratospheric, SO2, NO2, N2O, NOx, greenhouse gases, sulphur dioxide, carbon dioxide, ground-level ozone, water vapor, carbon monoxide, GHGs, nitrogen oxide
2	Ultraviolet radiation (topn = 20)	UV radiation, ultraviolet, UVR, UV, radiation, thermal radiation, UV-B, solar radiation, sunlight, irradiances, radiations, UVB, EUV, sunburn, PAR, radiant energy, photochemical reactions, photons, infrared, re-emits
3	Earth (topn = 40)	earth, planet, Earth's, Earthâ\x80\x99 s, surface_of_the_Earth, Moon, earth's, Earth's_surface, earth\x92 s, Earth\\, Sun, sun, Venus, Mars, planet's, earthâ\x80\x99 s, moon, Saturn, magnetosphere, Jupiter, surface, atmosphere, planetâ\x80\x99 s, globe, Uranus, planets, ground, orbit, crust, ocean_floor, sphere, Pluto, above, objects, object, photosphere, Earth's, body, icy, earth\\
4	Plant (topn = 40)	plants, animal, bacterial, tissue, host, microorganism, microbial, microorganisms, crop, biomass, woody, fungus, legume, seed, tissues, fish, weeds, fruit, aquatic, organism, caterpillar, fungal, seeds, parasite, tree, fungi, microbes, bacteria, forest, food_chain, organisms, rice, insect, plant's, microalgae, pathogen, micro-organisms, weed, photosynthetic, yeast
5	Animal (topn = 40)	organism, fish, animals, plant, insect, aquatic, shellfish, excreta, bird, food, faeces, wild, adult, host, edible, living, dung, feces, human, insects, wildlife, carcasses, pets, humans, seaweed, meat, larva, intestines, Elodea, sedentary, mammal, exotic, fishes, herbivores, chicken, eating, reptile, livestock, microorganism, decomposers
6	Human (topn = 40)	humans, humankind, health, natural, environmental, humanity, social, economic, mankind, biological, society, natural_processes, welfare, animal, cultural, Human, climate_change, man, ecological, our, anthropogenic, their, fetus, wildlife, global_warming, life, grave, detrimental, fear, living, technological, well-being, respiratory, peopleâ\x80\x99 s, distress, person's, agricultural, adverse, natural_resources, lifestyle
7	Absorb (topn = 20)	emit, penetrate, photosynthesize, lose, dissipate, evaporate, consume, attenuate, dissolve, redistribute, retain, radiate, absorbs, adsorb, vaporize, enter, disperse, deplete, sequester, trap
8	Create (topn = 20)	produce, generate, develop, make, deliver, provide, induce, impart, give, acquire, accommodate, add, lead, bring, incorporate, drive, offer, handle, render, allow
9	Destroy (topn = 20)	kill, disrupt, impair, disturb, degrade, suppress, threaten, overwhelm, deplete, eliminate, stimulate, infect, displace, regenerate, undermine, pollute, invade, render, resist, eradicate

6 Conclusions and Future Work

In this paper, we have presented a methodology for predicate-argument analysis with two objectives: improving conceptual relation expressiveness and designing a phraseology module. We have shown that semi-automatic and automatic approaches can be combined and reinforce one another. Results have only been provided for a small case study. In the near future, we will apply a combined top-down and bottom-up methodology to establish all basic semantic categories in the environmental domain. The top-down method will consist of a manual classification based on the definitions, conceptual networks, and other information contained in EcoLexicon. This will result in a domain-specific ontology similar to that of CPA semantic types, which is used in the Pattern Dictionary of English Verbs (PDEV) [19] for general language. We will again apply automatic clustering techniques to validate this manual categorization. The bottom-up method will consists of extracting all the verbs from the EcoLexicon corpus with TermoStat[4] (Drouin 2003) and then classifying them into different paradigms based on the concepts they relate and the basic conceptual relations they express, along the lines of the methodology described in this paper, to extract all the necessary information to populate our phraseology module.

Acknowledgments. 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.

References

1. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The sketch engine. In: Proceedings of the 11th EURALEX International Congress, pp. 105–116. EURALEX, Lorient (2004)
2. Faber, P., Mairal, R.: Constructing a Lexicon of English Verbs. Mouton de Gruyter, Berlin/New York (1999)
3. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop (2013)
4. Faber, P., León-Araúz, P., Reimerink, A.: Representing environmental knowledge in EcoLexicon. In: Bárcena, E., Read, T., Arús, J. (eds.) Languages for Specific Purposes in the Digital Era. EL, vol. 19, pp. 267–301. Springer, Cham (2014). doi: 10.1007/978-3-319-02222-2_13
5. Hausmann, F.J.: Le dictionnaire de collocations. In: Hausmann, F.J., Reichmann, O., Wiegandand, H.E., Zgusta, L. (eds.) Wörterbücher/Dictionaries/Dictionnaires — Ein internationals Handbuch zur Lexikographie/An International Enyclopedia of Lexicography/Enyclopédieinternationale de lexicographie, pp. 1010–1019. Walter de Gruyter, Berlin/New York (1989)
6. Gläser, R.: Relations between Phraseology and terminology with special reference to English. ALFA **7**(8), 41–60 (1994/1995)
7. Rundell, M.: Defining elegance. In: de Schryver, G.-M. (ed.) A Way with Words: A Festschrift for Patrick Hanks, pp. 349–375. Menha Publishers, Kampala (2010)

[4] http://termostat.ling.umontreal.ca/.

8. Bartsch, S.: Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence. Gunter Narr Verlag, Tübingen (2004)
9. Buendía Castro, M.: Verb dynamics. Terminology **18**(2), 149–166 (2012)
10. Buendía-Castro, M., Montero-Martínez, S., Faber, P.: Verb collocations and phraseology in ecolexicon. In: Kuiper, K. (ed.) Yearbook of Phraseology, pp. 57–94. De Gruyter Mouton, Berlin (2014)
11. L'Homme, M.C.: Predicative lexical units in terminogy. In: Gala, N., Rapp, R., Bel-Enguix, G. (eds.) Language Production, Cognition, and the Lexicon, pp. 75–93. Springer, Heidelberg (2015). doi:10.1007/978-3-319-08043-7_6
12. Alonso, A., Millon, C., Williams, G.: Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science. In: Proceedings of eLex 2011, pp. 12–22 (2011)
13. Fillmore, C.J.: Frames and the semantics of understanding. Quad. di Semant. **6**, 222–254 (1985)
14. Faber, P.: Frames as a framework for terminology. In: Kockaert, H.J., Steurs, F. (eds.) Handbook of Terminology, vol. 1, pp. 14–33. John Benjamins Publishing Company, Amsterdam (2015)
15. Faber, P., León-Araúz, P., Reimerink, A.: EcoLexicon: new features and challenges. In: Kernerman, I., Kosem Trojina, I., Krek, S., Trap-Jensen, L. (eds.) GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference, pp. 73–80. Portorož (2016)
16. Sahlgren, M.: The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Dissertation. Stockholm University (2006)
17. Brown, P.F., De Souza, P.V., Mercer, R.L., Della Pietra, V.J., Lai, J.C.: Class-based n-gram models of natural language. Comput. Linguist. **18**(4), 467–479 (1992)
18. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. J. Am. Stat. Assoc. **101**, 1566–1581 (2006)
19. Hanks, P.: Mapping meaning onto use: a pattern dictionary of english verbs. In: AACL 2008, Utah (2008)