Development of a Conceptual Lexicon with ontological techniques[a]

Katerina Tzortzi[1], Stella Markantonatou [2]
[1]*National and Kapodistrian University of Athens*
[2]*Institute for Language and Speech Processing/Athena RC, Athens, Greece*
tzortzi_katerina@yahoo.gr, marks@ilsp.athena-innovation.gr

Abstract

We report on a step-wise method for populating a computational conceptually organised lexicon of everyday language with concepts and words. We have worked with 'EKFRASI', a conceptually organised lexicon of Modern Greek that is being developed as a lexical ontology at ILSP/ 'ATHENA' RC. Our method is basically corpus driven and is initialized with a small number of top-down selected concepts. We developed a system for annotating verb dependents in Modern Greek corpus data and used it to identify the types of event denoted by a set of verbs that was used as a 'seed'. The 'seed' verbs were selected because they were profoundly related with HEALTH ---our case study top concept. The identified event types as well as the corresponding types of participant provided new concepts and words. Finally, general semantic relations such as "x hyperonym of y" as well as relations pertaining to the particular semantic field such as "x afflicts y" were applied on the already collected material resulting in a taxonomy of 22 concepts and about 670 words.

Keywords: Conceptual Lexicon, lexical ontologies, top concept, corpus driven, event type, semantic relations.

1. Introduction

This paper proposes a methodology for populating a conceptually organized lexicon of Modern Greek, implemented with the technology of ontologies, with concepts and words. We report on a mixture of top-down and bottom-up procedures.

2. 'EKFRASI': a conceptually organized lexicon

'EKFRASI' is a computational lexicon of Modern Greek (MG) that aims to help the user to successfully look up words even when s/he has few clues, for instance, to look up the word 'surgery' starting from the word 'to injure'. It is developed at ILSP/RC ATHENA drawing on the technology of ontologies and it is encoded with the Protégé ontology-editing tool. Ontologies allow for the definition of relations among the words. These relations are often only implicit in a printed lexicon (Gangemi et al.,

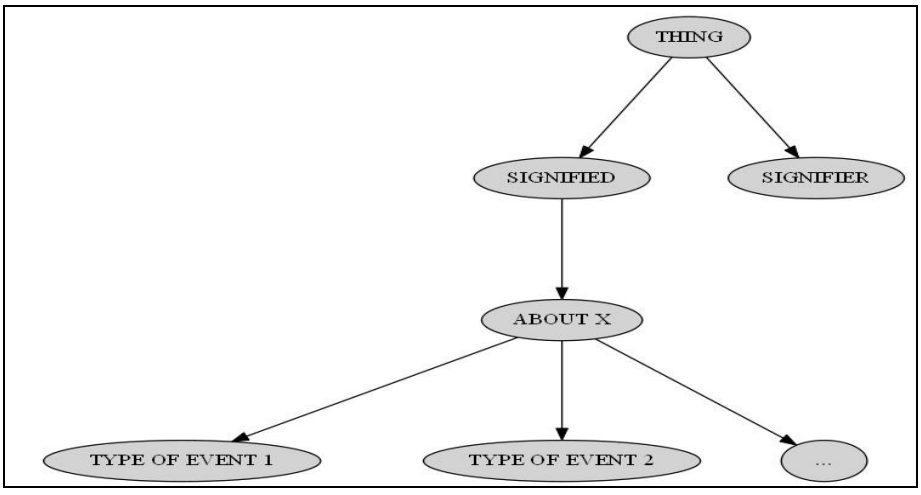
[a] The research leading to these results has received funding from: POLYTROPON (KRIPIS-GSRT, MIS: 448306)

2003). Words in 'EKFRASI' are organized according to their conceptual and lexical relations as well as their morphosyntactic properties (Markantonatou and Fotopoulou, 2007).
The ontology of 'EKFRASI' is based on the Saussurian distinction between the SIGNIFIER and the SIGNIFIED. The branch that is rooted at the class SIGNIFIER represents the morphological, syntactic and functional properties of words while the one that is rooted at the class SIGNIFIED represents word meaning. Words are instances of the SIGNIFIER and word meanings of the SIGNIFIED.
The inspiration for the development of 'EKFRASI' was drawn from Onomasticon (Vostantzoglou, 1962), a conceptually organized lexicon of MG in the line of Roget's Thesaurus (Trapalis et al., 2005). Hüllen (2004) argues that, in modern theoretical terms, Roget's "conceptual" organisation can be studied in the framework of semantic field theory. Like Roget's, Onomasticon encodes semantic fields in which words are organized in sets of synonyms that are reminiscent of the WordNet synsets (Fellbaum, 1998).

2.1 Structuring the space of concepts in 'EKFRASI'

Semantic fields in 'EKFRASI' are modeled as independent concepts that are subclasses of the class ABOUT X that, in turn, is a sub-class of SIGNIFIED (Pic.1). Immediate daughters of ABOUT X are general concepts such as ABOUT HEALTH, ABOUT ECONOMY, ABOUT EMOTIONS etc At this phase of EKFRASI development, we adopt general concepts from Onomasticon. However, as our work has already shown, we will not adhere to Onomasticon's general concepts (Section 3.3) rather we will be guided by our corpus data. The daughters of these very general concepts are types of event/situation defined on the basis of corpus work. We have used the notion of 'event/situation' that is also used by FrameNet (Baker et al.,1998) to define our 'types of event' on the basis of world knowledge and of corpus data. They are understood to encapsulate in their meaning sets of entity types (abstract or concrete) that stand in the relations prescribed by the specific event type and, possibly, other sub-events/situations.
Here, we take the general concept ABOUT HEALTH as a case study. We have adopted the concept from Onomasticon. Its daughter concepts defined on the basis of corpus work (Section 3.1) are ABOUT CURE EVENT, ABOUT UNHEALTHY SITUATION and ABOUT HEALTHY SITUATION (Pic 2).

Picture 1. The SIGNIFIED in 'EKFRASI'



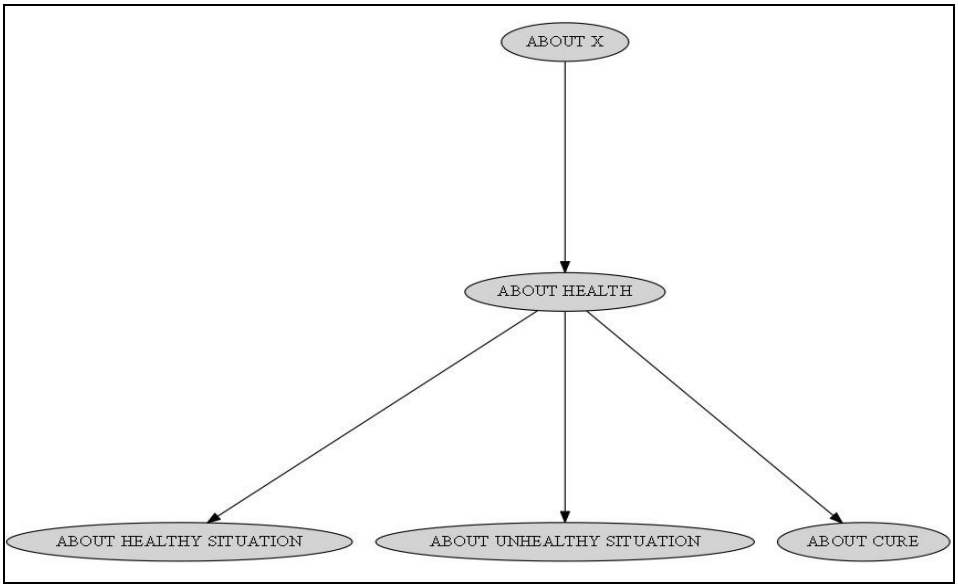
Event/situation types may or may not have daughter concepts. Their daughters may be other events/situations or collections of types of entity. Each time, these decisions are based on corpus data and on world knowledge. For instance, our data showed that a 'hospitalisation event' could be defined. We considered it **part of the cure event** because of world knowledge and because of our data (3.3): according to the corpus retrieved data, the semantic agent of the cure event 'points' to a general type of entity, that of 'healer'. World knowledge says that the 'healer' can sometimes belong to the medical stuff. When this happens, corpus data verify that the procedure of cure often takes place in a medical establishment, which is typically a hospital, and in that case we can define the 'hospitalization event' as part of the 'cure event' (3). Furthermore, our corpus data verified that the cure event typically involves in addition to the medical personnel, medical establishments, medical tools etc and patients. These types of entity can be viewed as **integral components** of the cure event, therefore we defined the classes ABOUT MEDICINE and ABOUT PATIENTS as subclasses of ABOUT CURE EVENT. In a similar way, patients were shown to play an integral role in ABOUT UNHEALTHY CONDITION SITUATION therefore, the instances of ABOUT PATIENTS instantiate **both** the ABOUT CURE EVENT (3) and the ABOUT UNHEALTHY CONDITION SITUATION.

Drawing on the discussion so far we describe the relation we have employed to structure the space of concepts in EKFRASI as an inclusion relation (Chaffin and Hermann 1984, Winston et al.,1987) that we call the 'IS ABOUT' relation. We define the IS ABOUT relation as follows: Class B IS ABOUT Class A iff Class B is an integral component or a subpart of Class A. Therefore, IS ABOUT is a relationship of relevancy rather than a hyperonymy-hyponymy one and it is a reflexive, antisymmetric and transitive relation.

The 'instance_of' relation that is available to ontologies encodes the IS A relation in EKFRASI. For example, "dermatologist" is a type of entity that belongs to the group of types of entity called "doctors". We have considered "dermatologist" an instance rather than a sub-class of the class ABOUT DOCTORS. In order to define a class such as ABOUT DERMATOLOGISTS we should have corpus evidence that a type of event exists where 'dermatologists' play an integral role. However, our data did not provide us with such types of event. Instead, all events involving 'dermatologists' belonged to the set of medical event types that is encoded with the class ABOUT MEDICINE. The class ABOUT DOCTORS was defined as a subclass of ABOUT MEDICINE because doctors were shown by our corpus data to play an integral role in medical events. So, in EKFRASI "dermatologist" is an instance of the class ABOUT DOCTORS (Pic. 3); due to the IS ABOUT relation, the concept "dermatologist" eventually belongs to both the concepts ABOUT HEALTH and ABOUT CURE but it is not related to either of them with the IS A relation. We further structure the space of instance concepts with the 'is_hyperonym_of' property. For example, given the instance concepts 'χειρουργός' (surgeon) and 'καρδιοχειρουργός' (cardiac surgeon) we define the relation 'χειρουργός is_hyperonym_of καρδιοχειρουργός'.
The proposed global relation ABOUT X structures meanings encoded in language more naturally than IS A that is known to structure only subareas of the language (Ježek and Hanks, 2010), mainly sets of words denoting entities of the physical world. IS A is a rather specialised relation. In language more general relations of 'relevancy' seem to play an organising role. This is how we interpret the basic idea behind Fillmore's Semantic Frames (1982): each word invokes a relevant semantic frame, ie.

a type of event/situation which defines the word's meaning. For example the concept *χειρουργός* ('surgeon') is not a kind of the concept 'medical event' but it immediately invokes it together with the lexical wealth it is related to.

Picture 2. 'ABOUT HEALTH'



3. Populating 'EKFRASI' with concepts and words

We report on a mixture of a top-down and a bottom-up procedure for populating 'EKFRASI' with concepts and words. We take the concept ABOUT HEALTH as a case study. Its definition was a top-down procedure; it was a top concept in Onomasticon that we adopted on the basis of common sense. From this point on, the procedure was "bottom-up", namely corpus-driven and resulted in the definition of 2 additional top concepts, 22 "lower" concepts and the encoding of 662 words.

3.1 Data collection

A small number of verbs undoubtedly relevant to the concept ABOUT HEALTH were studied: *χειρουργώ* ('to operate on'), *εγχειρίζω* ('to operate on', formal), *γιατρεύω* ('to cure' colloquial), *θεραπεύω* ('to cure', formal), *τραυματίζω* ('to injure') and *εξετάζω* ('to examine'). Our goal was to check our methodology that consisted in using as a "seed" (or as 'axioms') a small set of verbs whose meaning is undoubtedly related to a specific semantic field in order to delineate and populate a specific semantic field with concepts.

Data were collected from the Hellenic National Corpus (http://hnc.ilsp.gr), a balanced corpus of Modern Greek that allows for lemma searches. We retrieved about 1790 sentences in which the verbs occurred with a sense belonging to the ABOUT HEALTH semantic field. As we will describe below, we also used the Web and questionnaires for data retrieval.

3.2 Data annotation

A system of semantic and grammatical annotation of the retrieved sentences was developed (Tzortzi, 2014) drawing on the semantic role literature (Fillmore 1968, Dowty 1991, Wechsler 1995). (1) represents a semantically annotated sentence.

(1) [ΕΥ: αυτή = [δράστης] θεραπεύει ασθενείς [πάσχων] από χολέρα [πάθηση]
'She [agent] cures patients [patient] from cholera [illness]'

Next, we added a layer of grammatical annotation to the already semantically annotated sentences. For every syntactic constituent of the verb, we declared its grammatical function and its thematic role. Table 1 represents the syntactic and semantic annotation of (1).

Table 1. Example of grammatical annotation

word	grammatical annotation
αυτή 'she'	ΟΦον [δράστης] 'NPnom [agent]'
θεραπεύει 'cures'	Ρήμα 'Verb'
ασθενείς 'patients'	ΟΦαιτ [πάσχων] 'NPacc [patient]'
από χολέρα 'from cholera'	ΠΦ [πάθηση] PP [illness]

This procedure allowed us to describe in detail the semantic and syntactic properties of the verbs studied.

3.3 Structuring the semantic field ABOUT HEALTH

As said, a small number of verbs were taken as a 'seed'. The delineation and population of the ABOUT HEALTH concept drew on the results of the study of the 'seed' (3.2) in the way described in this section.

Following FrameNet (Baker et al., 1998), the verbs studied were first classified according to the event type they defined. Two event types were eventually obtained, the 'cure event' (2) and the 'affliction event' (3).

(2) CURE EVENT (merged): A**n agent (volitional)** applies some medical procedure/cures a **patient (an entity)** on some **part of the entity's body** because of some **illness** in some specified **manner/**with particular **means** at some specified **place** and **time**.

(3) AFFLICTION EVENT: An **agent (volitional)/**a **cause** afflicts a **patient (an entity)** on some **part of the entity's body**.

The two events were different from each other. In (2) 'agent' was entailed to be an intentional entity while in (3) the event was caused by an agent (intentional entity) or a cause (such as an illness).

We have developed a system of semantic annotation (Tzortzi, 2014) influenced by Dowty's Proto-roles (1991). Our system is more analytical in certain aspects. Thus, while Dowty collapses the entailements of intentionality and causation under the same Proto-role, we define two different thematic roles: intentionality is related with the **agent** role (2) while the **cause** role (3) is not necessarily related with intentionality. Thus, in (2) and (3) an **agent** is always considered to act intentionally, while Dowty's Ptoto-agent property: 'causing an event or change of state in another participant' (Dowty 1991:572) is described with an independent thematic role, the **cause**.

(2) and (3) led to the definition of two concepts: ABOUT CURE and ABOUT UNHEALTHY CONDITION. We further structured our lexical material with concepts relevant with ABOUT CURE and ABOUT UNHEALTHY CONDITION. For example, we decided to put ABOUT MEDICINE (the set of medical event types) below ABOUT CURE because the 'cure event' (2) resulted from the first version of 'cure event' (5) and the merging of the 'medical operation event' (4) which is one of the events that belong to the 'medical event' type in MG.

(4) MEDICAL OPERATION EVENT: A **surgeon (agent)** operates on **a patient (an entity)** some **part of the entity's body** because of some **illness** in some specified **manner** at some specified **place** and **time**.

(5) CURE EVENT (simple): A **healer (agent)** cures **a patient (an entity)** on some **part of the entity's body** because of an **illness** with particular **means** at some specified **place** and **time.**

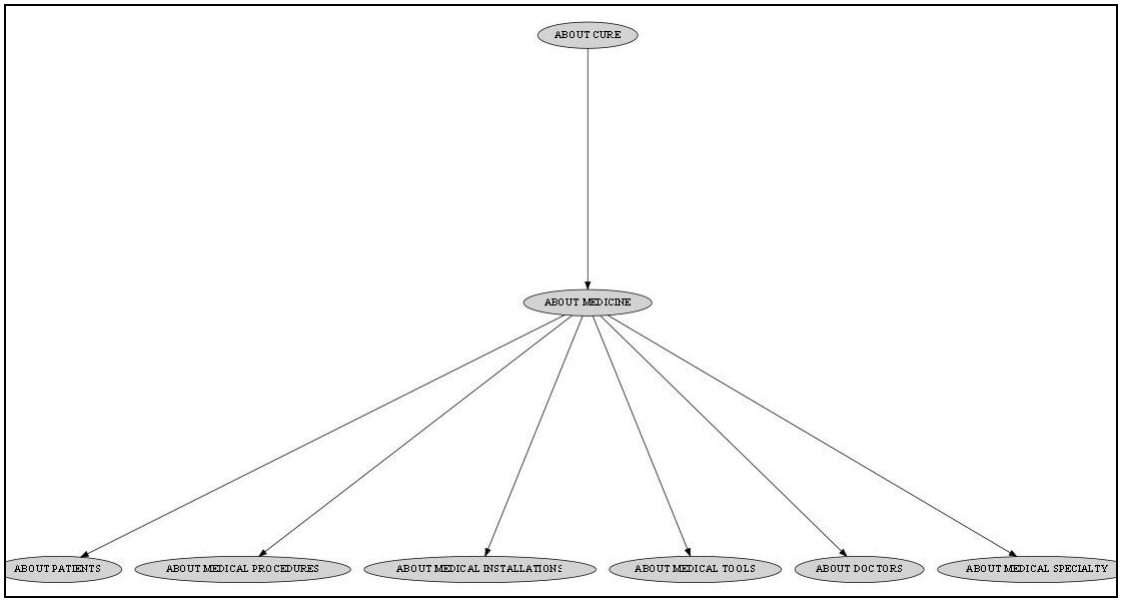
As we observe, the 'medical operation event' (4) and the first version of the 'cure event' can be described as subevents of the same more general event type (2) with some small differences. The basic difference is the type of the agent in each event type. As the corpus-retrieved data suggested **agent** in (4) 'points' to the types of entity 'surgeon' and 'medical stuff' while in (5) **agent,** besides the type of entity 'medical stuff', also 'points' to a more general type of entity, that of the 'healer' and certainly there are 'healers' who do not belong to the medical stuff. This difference led us to define ABOUT MEDICINE as a concept that is related to the concept of ABOUT HEALTH not directly, but through the concept of ABOUT CURE (Pic.3). Apart from this difference the two events described above share several main types of integral component (entity, part of the entity's body, illness, place and time), therefore the event types (4) and (5) can be 'merged' in the same event (2) (the one is a subclass of the other). Certainly, 'the integral components' hold a complement role in the VPs denoting the relevant events.

In order to delineate the semantic fields and perhaps discover more relevant ones, we used the verbal complements combined with their semantic function as it was defined with the procedure of data annotation of the corpus-retrieved examples (3.2). For example, the participants in the medical operation event (4) 'point' to several senses:

- **agent** 'points' to the types of entity '*surgeon'* and '*medical stuff*'
- **patient** (used here in its linguistic meaning) 'points' to the type of entity '*patient*' (with the every day meaning of the word and not the linguistic one)
- **the part of the entity's body** 'points' to the type of entity '*body part*'
- **illness** 'points' to the type of entity of '*illness*'
- **manner** 'points' to the type of event '*medical procedure*' and the type of entity '*medical tool*'
- **place** 'points' to the type of entity '*medical installation*'

Each of the above types was defined as a sub-class of the ABOUT MEDICINE class that is, as already said, a subclass of ABOUT CURE (Pic.3). On the other hand, the concept ABOUT BODY (and the sub-concept ABOUT BODY PARTS) was defined as a top concept (in the same way that ABOUT HEALTH was) because, although the semantic field about "human body" is directly related to ABOUT HEALTH it was clear that it would also be directly related to other potentially top concepts such as ABOUT EMOTIONS or ABOUT FOOD, thus playing a central and independent role in language. This is an example of delineation of a semantic field as well as of bottom-up definition of top concepts (as opposed to the top-down one that was adopted in the set-off phase of 'EKFRASI' development).

Picture 3. 'ABOUT CURE'



3.4 Populating the semantic field ABOUT HEALTH with words

Next, semantic relations among the meanings/words already selected and other, new ones were used in order to populate our lexicon with words. Relations 1-6 were identified in a top-down fashion while relation 7 was a result of our increasing familiarity with the data:

1. **Synonymy**: *θεραπεία* ('cure',N) - *γιατρειά* ('cure',N)
2. **Antonymy**: *ασθένεια* (illness) - *υγεία* (health); similar and other data led to the definition of a third branch of the taxonomy, ABOUT HEALTHY SITUATION (Pic. 2).

3. **Hyponymy-Hyperonymy:** *γιατρός* ('doctor') hyperonym of *χειρουργός* ('surgeon'), *καρδιολόγος* ('cardiologist') ect. As explained in Section (2.1), in 'EKFRASI', the IS A relation is encoded as a local property of meanings and not as the ordering relation that defines the taxonomy. In fact, it is a relation that can only be applied among instances as in the case of medical specialties; each medical specialty is related with the hyponymy-hyperonymy relation with the concept ABOUT DOCTORS and they are all encoded as instances of this concept.
4. **is the action of** (relating nouns with verbs): *θεραπεύω* ('to cure') - *θεραπεία* (cure,N). "action" is used here to denote actions, activities, facts and dynamic procedures. Of them, actions and activities involve an agent while facts and dynamic procedures do not (Lyons, 1977).
5. **is the quality of** (relating nouns with adjectives that mostly declare a quality (Lyons 1977)): *ασθένεια* (illness) - *ασθενής* ('patient'). Inverse relation: *ασθενής* ('patient') **has the quality of** *ασθένεια* ('illness').
6. **is the cause of** (relating a cause with its effect): *νοσογόνος* ('morbid') - *αθένεια* ('illness'). Inverse relation: *ασθένεια* ('illness') is **the effect** of *νοσογόνος* ('morbid').
7. **afflicts entity** (relating a situation with an entity that suffers the result of this situation): *ζαλάδα* ('dizziness') - *ζαλισμένος* ('dizzy'). Inverse relation: *ζαλισμένος* ('dizzy') is **the entity afflicted** by *ζαλάδα* ('dizziness').

4. Conclusions

We have described the procedure we adopted for structuring an electronic conceptual lexicon of Modern Greek using the technology of ontologies. We proposed a methodology for populating such a lexicon with concepts and words based on corpus retrieved data. We showed that using a small number of words as a 'seed' we can define relevant concepts. We also proposed a number of semantic relations that seem to apply throughout the semantic field we have studied. Of course our methodology needs to be evaluated against more data. We expect that a lot will be learned from the study and encoding of a larger number of diverse semantic fields. At this point we would like to thank the anonymous TOTh 2014 reviewers of our paper. Their comments were excellent food for thought.

References

PROTÉGÉ http://protege.stanford.edu/download/download.html)

Baker, C., Fillmore, C., and Lowe, J. 1998. "The Berkeley FrameNet Project". In *Proceedings of the 17th International Conference on Computational linguistics*. Montreal, Canada.

Chaffin, R., and Hermann, D.J. 1984. "The similarity and diversity of semantic relations". *Memory and Cognition*, 12:134-141.

Dowty, D. 1991. "Thematic Proto-Roles and Argument Selection." *Language,* Vol. 67, No. 3:547-619.

Gangemi, A., Navigli, R., and Velardi, P. 2003. "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet." In *Proceedings of ODBASEO03*. Catania, Sicily, Italy.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Fillmore, C. 1968. "The Case for Case." In *Universals in Linguistic Theory*, edited by E. Bach and R. T. Harms, 1-88. Holt, Rinehart and Winston.

Hüllen, W. 2004. *A History of Roget's Thesaurus. Origins, Development and Design*. Oxford: Oxford University Press.

Ježek, E., and Hanks, P. (2010). What Lexical Sets Tell us about Conceptual Categories. *Corpus Linguistics and the Lexicon, Special Issue of Lexis, E-Journal in English Lexicology4*: 7–22.

Lyons, J. 1977. *Semantics*. Cambridge University Press.

Markantonatou, S., and Fotopoulou, A. 2007. "The tool 'Ekfrasi'." In *Proceedings of the 8th International Conference on Greek Linguistics, The Lexicography Workshop*. Ioannina, Greece.

Roget, P. 1995. *Roget's II: The New Thesaurus*, Third Edition. Available from: http://www.bartleby.com/62/11.html

Trapalis, G., Markantonatou, S., Alexopoulou, M., Fotopoulou, A., and Maistros, Y. 2005. "Developing a small scale semantically organized Lexical Database of Modern Greek." In *Proceedings of the 7th International Conference on Greek Linguistics*. York, UK.

Tzortzi, K. 2014. "The development of a semantic field in a Conceptual Lexicon". ("Το χτίσιμο ενός σημασιολογικού πεδίου σε ένα Εννοιολογικό Λεξικό") Master's thesis, National and Kapodistrian University of Athens and National Technical University of Athens.

Vostantzoglou, T. (Βοσταντζόγλου, Θ.) 1962. *Αντιλεξικόν ή Ονομαστικόν της Νεοελληνικής*. Αθήνα: Δομή.

Wechsler, S. 1995. *The Semantic Basis of Argument Structure*. Stanford,CA: CSLI Publications.

Winston, M., Chaffin, R., and Hermann, D. 1987. "A Taxonomy of Part-Whole Relations". *Cognitive Science*, 11(4):417-444.