The involvement of native speakers in the lexicographic phase of data acquisition[1]

Stella Markantonatou, Katerina Tzortzi

Institute for Language and Speech Processing/Athena RIC, Athens, Greece

marks@ilsp.athena-innovation.gr, tzortzi_katerina@yahoo.gr

1. Introduction

An issue that perhaps should be taken into account in discussing lexicographical workflow is what happens if for one or another reason work stops and is resumed later. In such a case, some (or all) of the phases of the lexicographical process may be re-evaluated and, perhaps, re-conducted. This is the case of Ekfrasis, a conceptually organised (to be on-line) lexicon of Modern Greek (Markantonatou and Fotopoulou, 2007).

2. Lexicographical workflow

2.1 Preparation

The preparation phase of Ekfrasis took place during the second half of 2006 (project **Ekfrasis** http://www.ilsp.gr/en/infoprojects/meta?view=project&task=show&id=62 (04/2006-09/2007). A vast survey of user requirements was conducted with the use of a mock-up and of personalized questionnaires (Markantonatou et al., 2010). As a result:

1. The main functionalities of the lexicon were identified. The lexicon would allow the user to retrieve rich information about a word even if s/he could not recall it. For instance, s/he would be able to retrieve the word 'tincture of iodine' starting from the word 'to injure' together with definition and rich grammatical and lexical information.
2. The technology of ontologies was adopted to structure and encode lexical knowledge.
3. The basic architecture of the lexicon was designed. The ontology would adopt the basic (saussurian) dichotomy, that of the signifier and the signified. The forms of the words and their grammatical properties would be instances of the signifier branch of the ontology while concepts would be instances of the signified branch of the ontology.
4. A basic lexicon user interface was described.

2.2 Data acquisition

The phase of data acquisition nearly overlapped with the phase of preparation mainly because the main reference corpus, the HNC http://hnc.ilsp.gr/, was already available on line. HNC, the largest available on-line balanced corpus of Modern Greek, is fully tagged and offers good search facilities. A largish collection of MWEs as well as established printed lexica were also available. Such a lexicon, namely the Ονομαστικον by Th. Vostantzoglou, a conceptually organised lexicon of Modern Greek in the line of Roget's Thesaurus, was the source of inspiration for Ekfrasis.

2.3 Computerisation

The phase of computerization was relatively lean. The Protégé ontology editor was selected. The relevant set up was made to ensure that each lexicographer could access the DB independently, however, a certain organization of the team had to be established in order to make sure that work

[1] This research was partially supported by POLYTROPON (KRIPIS-GSRT, MIS: 448306)

European Union
European Regional Development Fund
HELLENIC REPUBLIC
Ministry of Education and Religious Affairs
General Secretariat for Research and Technology
HELLENIC REPUBLIC
MINISTRY FOR DEVELOPMENT & COMPETITIVENESS
COMPETITIVENESS AND ENTREPRENEURSHIP
OPC II
NSRF 2007-2013
programme for development
regions at the centre of development
Co - financed by the Hellenic Republic and the European Union - European Regional Development Fund,
in the context of the O.P. Competitiveness and Entrepreneurship (OPC II) and the R.O.P. Attica, R.O.P. Macedonia - Thrace

was not overwritten or lost as such facilities were not guaranteed automatically. It was soon obvious that Protégé did not offer the right interface for demanding lexicographic work but no alternatives could be offered at the time.

2.4 Data processing-first phase

The overall effort ended at a point that would be equivalent to the phase of data processing (Jan 2007 to Sept 2007). Frequency lists of the content words of HNC were obtained. Lexicographers worked on selected semantic fields drawing on these lists and on printed lexica. Also, more than a thousand MWEs were encoded. About 4000 lemmata were encoded in this way. However, this phase was rather experimental. Several questions presented themselves, such as:

1. The 'instance-of' relation of Protégé was used to encode relations such as the one between the concepts 'tincture of iodine' and 'anticeptic'. Beyond that, the precise structure of the signified branch remained an issue. Ekfrasis was meant to encode 'semantic fields' rather than IS A hierarchies. After all, the IS A relation structures only certain areas of the lexicon, particularly the ones that 'refer to' entities, typically nouns in Modern Greek. However, Ekfrasis aimed to express semantic relations holding among words that belong to all parts of speech and are not necessarily related to each other on the basis of their morphology. A rather loosely defined relation among concepts called 'ABOUT X' was chosen.
2. At this experimental stage, it became clear that a method should be developed for (i) deciding which words belong to a certain 'semantic field' (ii) identifying the semantic relations among these words that would ensure navigation through the 'semantic fields'.
3. Last, but not least, was the issue of the adequacy of the primary data resources. HNC is a corpus of rather moderate size (50M) while the Ονομαστικόν was last printed in 1964.

Ekfrasis run out of funding in September 2007. The issues presented above were left open to research. In 2013 the second phase of Ekfrasis development started. Meanwhile, the original group of lexicographers has withered out and new people have taken up, while different researchers have taken different paths in their effort to provide an answer to the problems identified.

2.5 Data processing-second phase

A phase of reevaluation of primary resources was necessary again because the issue was not investigated thoroughly in the first phase of Ekfrasis development (2006-2007) and, as it was already said, resources were of moderate size and/or relatively outdated.

Table 1. A picture of the data retrieved from HNC

Verb	Medical sense	HNC	Medical sense occurrences (MSO)	Active forms	% on MSO	Medio-passive forms	% on MSO
χειρουργώ	to operate on	198	176	62	35	114	65
εγχειρίζω	to operate on	62	29	7	24	22	76
εξετάζω	make a medical examination	2000	36	36	100	0	0
τραυματίζω	to injure	2000	1306	214	16	1092	84
γιατρεύω	to cure, colloq.	115	44	30	68	14	32
θεραπεύω	to cure, formal	513	191	88	46	103	54

We used the semantic field of HEALTH as a case study. First, we studied the medical event (we use the term "event" in the way it is used in FrameNet (Baker et al.,1998)). We set off by investigating the syntactic behavior of six verbs undoubtedly related with health: *χειρουργώ* ('to operate on'), *εγχειρίζω* ('to operate on', formal), *γιατρεύω* ('to cure' colloquial), *θεραπεύω* ('to cure', formal), *τραυματίζω* ('to injure') and *εξετάζω* ('to examine') (Tzortzi and Markantonatou 2014). We drew data from the HNC, the Web and the Questionnaire.

HNC returned about 4900 sentences out of which 1790 sentences featured 'medical' usages of the verbs in question (Table 1). The semantic and grammatical annotation (Tzortzi 2014) of the HNC data provided a detailed picture of the structures supported by these verbs. Table 2 offers the relevant data for the verb *χειρουργώ* ('to operate on'). Still, being native speakers of MG, we felt that there were structures missing. We checked web-retrieved examples and identified some additional structures (Table 3) but not all. So, we designed a questionnaire. We also used statistical methods to verify some of the assumptions we made with the corpus-retrieved data in mind.

We distributed the questionnaire to 75 native speakers of MG aged between 14 and 62 years old. We tried to make sure that the questionnaire would not put any bias on the speaker's intuitions. Each speaker was exposed to two examples only, each one featuring a different structure. The texts were chose to make sense on their own, all of them being very short narratives. We gave the informants no hint about what we were after hoping that, if there was something wrong with the verb usage, they would spot it intuitionally. The questionnaire showed that the HNC contained only 40% of the structures supported by *χειρουργώ* ('to operate on') (Table 4).

Table 2. Active structures of *χειρουργώ* ('to operate on') in the HNC

	Structures	**%**	**Usage example**	**English Translation**
1	Subj-*Agent* Verb Obj-*Entity*	66,2	…έτυχε παλιά να χειρουργήσω μερικούς από τους πιο στενούς μου φίλους	…it happened to me in the past to operate on some of my best friends
2	Subj-*Agent* Verb Obj-*Entity για*-PP-*Illness*	1,6	Ένα παιδάκι που χειρουργούμε για αμυγδαλές	A child which we operate on for tonsillitis
3	Subj-*Agent* Verb [Obj-*Bodypart* +POSS-*Entity*]	3,2	Ο ίδιος είχε χειρουργήσει τη μύτη του Φρόυντ δύο φορές	He himself had operated twice on Froyd's nose
4	Subj-*Agent* Verb Obj-*Bodypart*	1,6	Χειρουργεί μόνο θυρεοειδείς.	He operates on thyroids only.
5	Subj-*Agent* Verb	27,4	Ενώ αυτός που χειρουργεί δεν μπορεί να χειρουργεί εν τη απουσία του....	While the one who operates on can not operate on while he is absent…

Table 3. Active structures of *χειρουργώ* found in the Web

	Structure (active forms of 'χειρουργώ')	**Usage example**	**English Translation**
6	Subj-*Agent* Verb Obj-*Entity* σε-PP-*Bodypart*	…προκειμένου να συναντήσει τον γιατρό που τον χειρούργησε στο γόνατο	in order to meet the doctor who operated on his knee
7	Subj-*Agent* Verb Obj-*Entity* [για-PP-*Illness*+ POSS- *Bodypart*]	ο γιατρός που τον χειρούργησε για καρκίνο του παχέος εντέρου…	the doctor who operated on him for lower intestine's cancer
8	Subj-*Agent* Dative-*Entity* Verb Obj-*Bodypart*	εξετάστηκε εκ νέου από τον καθηγητή-γιατρό που του χειρούργησε το αριστερό γόνατο.	he was examined again from the doctor who operated on his left knee
9	Subj-*Agent* Verb Obj-*Illness*	Υπάρχει μια ομάδα εξειδικευμένων χειρουργών που μπορεί να χειρουργήσει όγκους κοντά στο σφιγκτήρα	There is a team of specialized surgeons who can operate on tumors near the clamp
10	Subj-*Agent* Verb [Obj-*Illness* +POSS-*Entity*]	Ποιος χειρούργησε το λίπωμα της Μενεγάκη;	Who operated on Menegaki's lipoma?

During the data processing phase, we tested assumptions based on corpus frequency data using statistical methods (Gries, 2012) and the R programming language (http://cran.r-project. org/bin/windows/base/). Statistical tests showed that conclusions should not be based on corpora frequencies only. For example, HNC frequency tables suggested that the verb *θεραπεύω* ('to cure',

formal) prefers the passive morphology while its synonym *γιατρεύω* ('to cure', colloquial) the active one. However, the statistical tests showed that both verbs prefer the active form.

In this paper we brought evidence that corpora and web retrieved data do not provide the whole picture, rather they provide the most frequently used structures. Here comes the question of how much information should be included in an electronic lexicon. Since there are no space limitations in an electronic lexicon and native speakers tend to look up the rare usages of the words (de Schyver et al.,2006), we decided to include all the possible usages of the verbs in our lexicon---and this is why user involvement was shown to be important for the phase of data acquisition. This leads us to re-evaluate the reference data we use to develop Ekfrasis: now, we know that we will provide the user with the most frequent usages and that, if we want to encode less frequent usages some crowdsourcing method should be devised, given the resource situation for Modern Greek.

Table 4. Additional usages of Active forms of 'χειρουργώ' ('to operate on') in the questionnaire

	Structures	**%**	**Usage example**	**English translation**
11	Subj-*Agent* Verb Obj-*Entity* σε-PP- *Bodypart* για-PP-*Illness*	100	Ένα παιδάκι που χειρουργούμε στον εγκέφαλο για καρκίνο	A little child that we operate on the brain for cancer
12	Subj-*Agent* Verb Obj-*Entity* [για-PP-*Illness* + σε-PP- *Bodypart*]	100	Οι ασθενείς τους οποίους χειρουργούμε για καρκίνο στον πνεύμονα μένουν στην απομόνωση	Patients who we operate on the lung for cancer
13	Subj-*Agent* Dative-*Entity* Verb Obj- *Bodypart* για-PP-*Illness*	93	Συνάντησα το γιατρό που μου είχε χειρουργήσει το στομάχι για έλκος.	I met the doctor who had operated on my stomach for ulcer
14	Subj-*Agent* Dative-*Entity* Verb Obj-*Illness*	100	Της χειρούργησε το κάταγμα ο Νικολακάκης	Nikolakakis had operated on her fracture
15	Subj-*Agent* Verb Obj-*Bodypart* για-PP-*Illness*	93	Το πρωί χειρούργησα ένα στομάχι για καρκίνο	In the morning I operated on a stomach for cancer
16	Subj-*Agent* Verb [Obj-*Bodypart* +POSS- *Entity*] για-PP-*Illness*	93	Τώρα τελευταία χειρούργησε το μάτι του Κ. για καταρράκτη	Lately he operated on K' eye for cataract

3. References

Βοσταντζόγλου, Θ. 1962. *Αντιλεξικόν ή Ονομαστικόν της Νεοελληνικής*. Αθήνα: Δομή.

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

De Schryver, G-M., Joffe, D., Joffe P., and Hillewaert, S. 2006. Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos 16* (AFRILEX-reeks/series 16: 2006): 67–83.

Gries, S. To appear. Frequency tables, effect sizes, and explorations. In Dylan Glynn & Justyna Robinson (eds.), *Polysemy and synonymy: corpus methods and applications in Cognitive Linguistics*. Amsterdam & Philadelphia: John Benjamins.

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

Markantonatou, S., Fotopoulou, A., Mini, M. and Alexopoulou, M. (2010). In search of the right word. In *Proceedings of Cogalex-2: Cognitive Aspects of the Lexicon, 2nd SIGLEX endorsed Workshop*. Beijing

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.

Tzortzi and Markantonatou 2014. The 'medical event' in a conceptually organized lexicon. In *Proceedings of the 11th International Conference on Greek Linguistics*. Rhodes island, Greece.