

Sharing the Knowledge of Lexicographers: Methodology for the Extraction of Lexicographic Abilities

Sophie Comeau

OLST—Département de linguistique et de traduction
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (QC) H3C 3J7
tel. (1-514) 343-6111 Ext. 32585
fax (1-514) 343-2284
Sophie.comeau@umontreal.ca

Abstract

In our paper, we will present the Lexitation project, which is, to our knowledge, the first attempt at extracting lexicographic abilities using experimental techniques. We will describe the methods and results of our initial set of experiments, that are based on the use of so-called Think Aloud Protocol (Ericsson & Simon, 1993). We will explain how experiments have been set up and how we are currently proceeding with the extraction and modeling of various types of knowledge and strategies used by lexicographers while performing lexicographic tasks. Finally, we will present possible applications of our work in the field of language teaching.

1 Introduction

Lexicography is a discipline that has been extensively documented (Zgusta, 1971; Béjoint, 2000; Landau, 2001, Mel'čuk 2006). Lately, a few authors have been discussing the relation between some of these writings, seen as “theoretical”, and the more “practical” field of dictionary-making (Fontenelle, 2008; Atkins & Rundell, 2008). However, there exists no systematic study of lexicographic abilities—i.e. declarative and procedural knowledge mastered by trained lexicographers—and no structured model of descriptive and methodological concepts involved in lexicography is available. Hence, the work we are presenting here targets the building of an ontological model of lexicographic abilities. It is part of a four-year project called Ontologization of lexicographic abilities for use in the fields of applied linguistics, nicknamed Lexitation, from which pedagogical applications will be derived. The project relies on the assumption that lexicographic abilities play a role in teaching and acquisition of lexical knowledge, and not only in lexicography per se. The originality of this research is that we extract lexicographic abilities using experimental techniques. Basically, we invite professional lexicographers to come and work before our eyes, and we record the work sessions. The recordings are later scrutinized to extract lexicographic abilities that we model into an ontology designed with the Protégé ontology editor.¹ Although our background is that of Explanatory Combinatorial Lexicology (Mel'čuk et al., 1995; Mel'čuk, 2006), we target the modeling of concepts and strategies used by lexicographers regardless of their specific theoretical background. We are still at the beginning of the project : five experiments took place so far. In the next sections, we will present results of the first stages of knowledge extraction from Experiment 3.

¹Protégé is a free ontology editor and knowledge-base framework, developed at Stanford University in collaboration with the University of Manchester.

2 Description of experimental approach

2.1 Experimental settings

The goal of our experiments is to extract the knowledge and know-how involved in lexicography. To do so, we need to observe some lexicographers at work. Hence, for each experiment, we invite a lexicographer to come to a laboratory where he or she performs a lexicographic activity that we record. Five experiments took place so far. The first two were used as tests; participants were students trained for Explanatory Combinatorial Lexicology. The other three participants are professional lexicographers, two for academic dictionaries and one for a private company. Two of them are from Québec and one is from France. We used a verbal protocol to gain information about their cognitive thoughts while performing a lexicographic task. Ericsson and Simon (1984) present concurrent (also referred to as Think Alouds) and retrospective (also referred to as Think Afters) verbal protocols as ways to generate data about cognitive processes. In concurrent verbal reports, participants are asked to report their thoughts during the performance of a task, whereas retrospective verbal reports rely on gathering information after a task is carried out. Although both techniques can be influenced by a motivational shift, that can occur whenever participants are informed they are being observed (Russo, Johnson & Stephens, 1989), they provide a great amount of data about cognitive and behavioral processes. Though the Think Aloud protocol can be a little invasive, it was chosen over the Think After protocol, because the latter is more likely to be influenced by forgetting and fabrication, as shown by Branch (2000) and Ericsson and Simon (1984). Moreover, we found out through our first experiments that the more the participants are experienced, the less they have difficulty in doing Think Alouds. This is consistent with the studies of Ericsson and Simon (1984) and of Branch (2000), who indicated that Think Alouds are less useful when participants are carrying out a new task or a task that involves a high cognitive load. Our participants all being experienced professional lexicographers (except for the test phase), they seemed not too disturbed by the Think Aloud protocol. Hence, we prefer to keep using that protocol, thus avoiding the problem of incomplete memories that can occur with the Think Afters. In the experiments, to monitor that procedure, we had a research assistant we called the “pilot” to sit next to the lexicographers and to encourage them to think out loud, by asking them questions such as “Why did you do that ?” or “Can you tell me what you are thinking at the very moment ?” etc. The pilot would also help them out whenever they had a technical problem. While performing the task, participants were audio and video taped by 3 cameras, and we had the computer screen recorded too. The lexicographer and the pilot would sit in a room, the workstation, while the “controllers” – i.e. other experimenters – would watch the recordings on TV screens from another room. They could see and hear what was going on in the workstation in real time, so they could address comments or instructions to the pilot through a microphone connected to the pilot’s headphone.

2.2 Lexicographic activities

The same lexicographic activity (see definition in 3.2) was used for the first three experiments. Participants were given the assignment to outline the polysemic structure of French vocable COMPTE, ‘to count’, that is, finding its different senses. The computer used was provided with French corpus Le Migou, build from French newspaper *Le Monde*, and participants could access the Web via a browser. They were told they could use any of the two corpora, or both, but that there was no obligation to. They were asked to work accordingly to their usual methods as far as possible. They were given up to an hour to fulfill the task, but we insisted it could be less if they didn't need as much time. They had to store each lexical unit they found in a separate file of a document, which model we had created previously with Filemaker software. They could create as many files as they needed. Each file form contained five fields for them to fill in:

1. LEXICAL UNIT, where they would write down the lexical unit’s name and a number if they wanted to (though they were not instructed to).

2. GRAMMATICAL CHARACTERISTICS, where they could write the lexical unit's part of speech, or any characteristic they found worthy of noticing.
3. CLUE, where they could write something that would allow them to distinguish the lexical unit from others. It could be a quasi-synonym, for example.
4. EXAMPLE, where they would write as an example a sentence containing the lexical unit. It could be their own example or it could be taken from corpus.
5. COMMENTS, where they could write any comment they found relevant.

That was the activity for Experiments 1 to 3. For Experiments 4 and 5, though, we asked the lexicographers to do, for an hour or so, the typical work they are trained to do, and that they actually do in their everyday job, i.e. writing or editing dictionary articles. They would therefore work on their own documents, using their own tools. Actually, we favor activities that simulate the usual lexicographic work of our participants, the objective being to benefit as much as possible from their participation and to extract as much lexicographic abilities as possible.

3 Methodology for data analysis

The extraction of lexicographic abilities that will be described here was made from data of Experiment 3 only. Data of other experiments are being currently analyzed.

3.1 Inductive techniques

Our approach to data analysis fits the category Ericsson & Simon (1984) call “meaningful analysis of verbalizations”, where there is no agreement, between the experimenter and the subject, upon specific signals. In this type of analysis, there are at least two ways to proceed. One way is to analyze the data in terms of their meanings, with a theory guiding the analysis, which limits the encoding to selected aspects and features and, when encoding the data, to map the verbalizations onto these categories of concepts and features. Our approach to extraction does not fit the above scheme. Our method was rather inductive, that is, we did not try to analyze the data along with a set of predetermined classes of lexicographic abilities. Indeed, we would rather scrutinize the “raw data” and try to identify what knowledge and strategies were used by the lexicographer along the experiment, and only then we would create corresponding abstract classes of lexicographic abilities. This procedure is perfectly legitimate, according to Ericsson & Simon (1989:6): “In less formal kinds of analysis, the encoding scheme is not defined formally and a priori, but the search for interpretations proceeds in parallel with the search for an appropriate model or theory. We recognize clearly the need for and value of such interactive processes in the search for theories in new domains”. Lexicography is not what we can call a “new domain” ; Dr. Johnson addressed his “Plan of a Dictionary” to Lord Chesterfield in 1747 and, since then, many authors have set theoretical foundations for the discipline (Zgusta, 1971; Béjoint, 2000; Landau, 2001; Mel’čuk et al., 1995; Mel’čuk, 2006, to name a few). However, there exist no structured model of descriptive and methodological concepts and abilities involved in lexicography. Therefore, when analyzing the data, we could not categorize each verbalization into one existing category. Although we plan to enrich our ontology with concepts taken from the literature once the analysis is over, we decided to start from the observation of the experiments to create categories of lexicographic abilities as they appear. As a result of this choice, we had to define our own methodology for data analysis and our own encoding scheme. We’ll explain how we proceeded in the next subsections.

3.2 Some definitions

Data analysis relies on three basic concepts. It is necessary to define them before to go any further with the description of the methodology:

LEXICOGRAPHIC ACTIVITY: The assignment given to the lexicographer for a given experiment. The activity is usually chosen and designed by the team of experimenters, but it can also be, like it was the case for Experiments 4 and 5, an activity taken from the lexicographer's daily duty. There is only one lexicographic activity per experiment.

LEXICOGRAPHIC TASK: Segment of the activity associated with a given time period in which the lexicographer accomplishes a specific task. What is important here is that the segmentation of the activity into tasks is a linear one; each task is given a number and its starting and ending time is identified. For Experiment 3, that lasted 55 minutes, we identified a total of 40 tasks, that means that tasks have a mean duration of nearly a minute and a half.

LEXICOGRAPHIC OPERATION: “minimal” action taken by the lexicographer inside a given task. Operations are not associated with given time sequences. We have to explain here that operations correspond to very short sequences that can overlap. They can also be inferred, that is, we can assume that the lexicographer does one particular operation even though he doesn't say so – If there are some evidences that allow us to believe so –. For it was difficult, if not impossible, at times, to locate operations in time, we decided to order them logically rather than chronologically.

The three concepts introduced here will be illustrated in next sections. Let us not forget that the analysis presented in this text was made with data from only one experiment; we expect various different types of tasks and operations to add up as other experiments are being analyzed.

3.3 Division of activity into tasks

After seeing recordings of Experiment 3, we first divided the footage in 40 great “scenes”, corresponding to the tasks accomplished by the lexicographer. These 40 tasks were mainly of three different types, according to the goal the lexicographer pursues in doing them. In our ontology, we created three abstract classes of tasks, corresponding to the three types of tasks we sorted out, and each of the 40 tasks is considered an instance of one of the three classes, which will be described here:

- **SEARCH FOR SENSE:** The lexicographer is searching for senses he haven't thought of yet, using the corpus or by introspection.
- **CONCEPTUALIZATION OF SENSE:** The lexicographer conceptualizes a lexical unit and writes a short lexicographic description in a file.
- **MODIFICATION OF DESCRIPTION:** The lexicographer operates a minor modification in one of his files.

We decided to include in the class **MODIFICATION OF DESCRIPTION** only the instances of tasks in which the modification does not imply a reorganization of the structure of the vocable as it is outlined by the lexicographer at the time. In general, a modification aims at improving a lexicographic description, by making it more complete or by conforming to dictionary-writing rules. For example, in task 18 of Experiment 3, the lexicographer adds an example taken from a corpus in his file **COMPTER-‘dénombrer’**, that already had an example made up by him. He adds it because an example taken from “real speech” would confer more legitimacy to his description. In this case, there is no conceptualization of a lexical unit, but a “superficial” modification in the lexicographic description.

3.4 Division of tasks into operations

At this stage of the analysis, we had in the ontology classes of lexicographic tasks and instances of these classes. Next step was to divide instances of tasks into simpler actions, called “operations”, and create abstract classes of operations, just like we did for the tasks. These two groups of classes are independents.

Classes of operations are not subclasses of tasks; the same type of operations can be performed in different types of tasks. Here are the classes of operations identified to date, and their hierarchy:

Note the following abbreviations:

LU means ‘lexical unit’; QSYN means ‘quasi-synonym’; ASSESS means ‘assessment’; EX means ‘example’; POS means ‘part of speech’; SPECIF means ‘specification’.

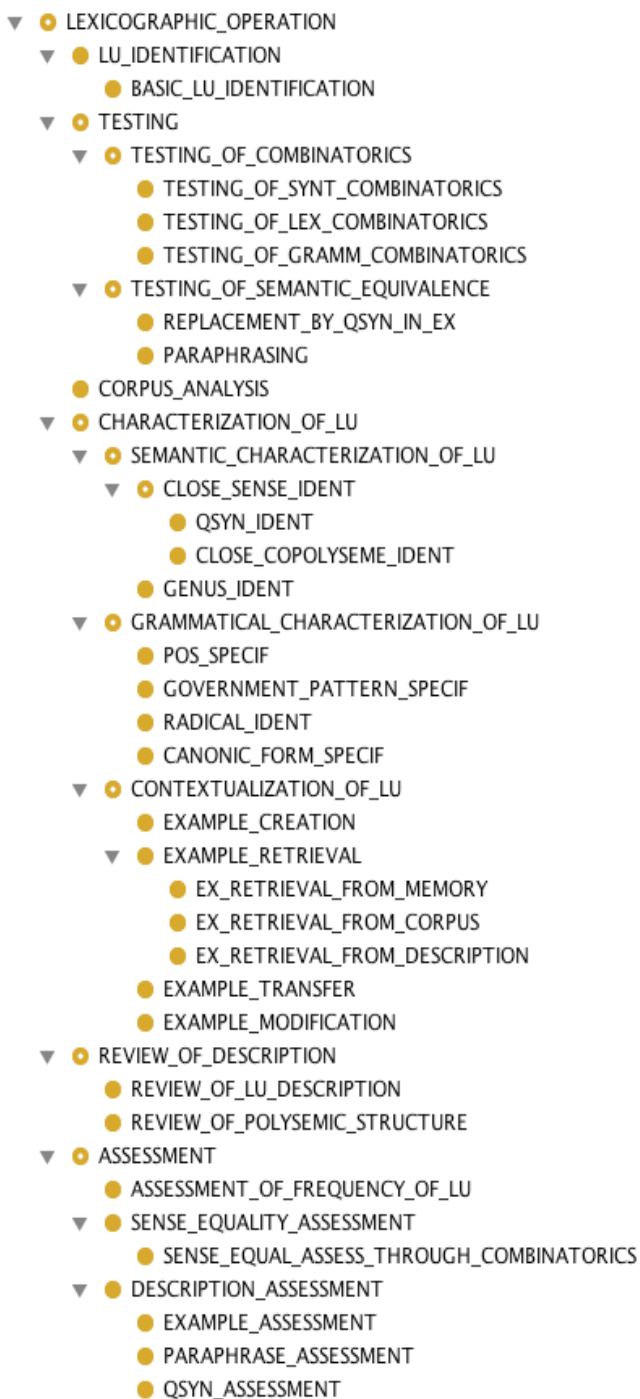


Figure 1. Hierarchy of lexicographic operations

Figure 1 displays all the classes and subclasses of operations that we observed in one or more tasks of the Experiment 3. In each specific task, the participant carries out some of these operations, but not necessarily in the order they are presented in the hierarchy. Classes of operations themselves are independent from one another. For example, the classes TESTING and LU_IDENTIFICATION are in a sister-sister relation rather than in a mother-daughter one, but BASIC_LU_IDENTIFICATION is a subclass (daughter) of the class LU_IDENTIFICATION. Other classes of operations will be added to these as data from other experiments are analyzed. As we mentioned before, operations are sometimes inferred, that is, we assume that the lexicographer carries them out even though he doesn't mention them. Of course, we try to infer as less as possible, but enough to obtain a complete and logic chain of operations. Most times, inferences rely on tangible evidences. For example, in task 4 of Experiment 3, lexicographer writes down the part of speech of the lexical unit he is thinking of, a verb, as he is talking about something else. Though he does not say "I am now identifying the part of speech of the lexical unit", we can infer he did and add POS_SPECIF to the list of operations of task 4. But since operations themselves are sometimes not tangibles, it is often impossible to identify exactly at what time they occur. Moreover, they are of very short duration and can often overlap. In this same task 4, for example, when the lexicographer is writing down the part of speech of the lexical unit, he is also uttering the hypothesis of this sense being distinct from another. The two operations take place simultaneously, and this is a pretty common case. Therefore, we decided not to identify time sequence of each operation, like we did for tasks. But what we can do is to order the operations according to an ordinal and logical order. Indeed, in the ontology, in each instance of operation, we identify the operations that led to that one. To make the methodology of the extraction more clear, let's see the transcript of a task taken from Experiment 3 and what classes of abilities we extracted from it.

4 Transcript of a task and its analysis

Here is the transcript of task 8 of Experiment 3. *Italic* is used to cite the lexicographer; translation follows, in parenthesis. Words in small capital letters are the names of the classes of operations we created.

1. « *Ça me fait penser, hier, je jouais avec ma nièce...elle a dit : " Ça, ça compte pas." »* (This reminds me, yesterday, I was playing with my niece...She said "that doesn't count").

= EX_RETRIEVAL_FROM_MEMORY

2. « *Donc... "C'est pas conforme au règlement." »* (So... "It doesn't go by the rule").

= QSYN_IDENTIFICATION

3. « *Ou... Il faut pas en tenir compte ... Il faut pas prendre en consideration... Ça ressemble un petit peu mais c'est pas "compter avec", là...»* (Or... "We shouldn't take it into account... We shouldn't take it into consideration... It's a bit similar, but it's not "compter avec").

= LU_IDENTIFICATION

First, in this operation, the lexicographer compares the sense of his retrieved example, «*Ça compte pas*», with the sense of another lexical unit of the same vocable, semantically close, that he tagged *COMPTER AVEC*- 'prendre en considération, tenir compte de'. (TO COUNT 'take into consideration'). Then, he states that the meaning of the lexical unit he retrieved is different from that of *COMPTER AVEC*, and he decides that this particular sense is worthy of the lexical unit status, so he creates a new file.

4. He identifies and writes down the part of speech of the lexical unit.

= POS_SPECIF

5. He writes down the quasi-synonym "être officiel".

= QSYN_IDENTIFICATION

6. He writes down another quasi-synonym, “réglementaire”.
= QSYN_IDENTIFICATION

7. He identifies another quasi-synonym, “homologué”.
= QSYN_IDENTIFICATION

8. « *Cette ronde de pratique ne compte pas.* » (This turn does not count).
= EXAMPLE_CREATION.

Here, he invents a sentence and writes it as an example.

9. « *Ça, ça compte pas.* » (That does not count).
= EX_RETRIEVAL_FROM_MEMORY

Here he recalls the same example he heard the day before, and he writes it down.

10. « *C'est peut-être un peu trop relâché.* » (This is maybe too informal).
= EXAMPLE_ASSESSMENT.

11. « *Je suis pas entièrement satisfait de ma description...Il y aurait peut-être quelque chose de plus juste.* » (I'm not entirely satisfied with my description...There could be something more accurate).
= QSYN_ASSESSMENT.

Here, he stares at his file, and when asked what he is thinking about, he says he is not entirely satisfied with his description, pointing at the quasi-synonym. Finally, he passes on to something else, saying that the quasi-synonyms are mainly there to help distinguish one lexical unit from the others.

We'll see now how classes of abilities are organized and encoded in the ontology.

5 Encoding of data

The encoding of data was done in an ontology designed with Protégé ontology editor. An ontology is a formal explicit description of concepts (classes) in a domain of discourse – in this case, lexicography –, properties of each concept describing various features and attributes of the concept (slots), and a set of individual instances of classes. Classes describe concepts in the domain, and slots describe properties of classes and instances (Noy & McGuinness, 2001). In our ontology, concepts correspond to the classes of lexicographic tasks and operations (as shown in Figure 1). Tasks and operations performed during the experiments are instances of these classes. In the ontology, there is a window for each instance of task or operation, in which various slots display information about the instance itself. For example, in each instance of task, we display its starting and ending time, the video of the task, and all the lexicological concepts used by the lexicographer during the task.

6 Pedagogical applications

As we said before, Lexitation project is in its beginnings. We are planning other experiments, some of which with English lexicographers. Experiments 4 and 5 will soon be analyzed, according to the method we exposed in this text. That being said, there is one question left to answer: «What purposes will the ontology serve?» We think it could be very useful, among other things, in language teaching, especially for teacher training. In the literature about language teaching, it is a pretty common place to say that lexicon is not given enough attention and that more work should be done in this field. In fact, authors from Québec (Simard, 1994), Switzerland (De Pietro, 2003) and France (Grossmann et al., 2005) deplore the lack of a systematic teaching of lexicon that would accompany the natural and spontaneous interventions about lexical features that teachers make on a daily basis as reading or other types of

activities are done in class. However, we think that the potential of our project lies right there, in the interventions about lexicon that teachers have to perform “on the fly”. These interventions are maybe not sufficient, but all teachers will agree that they are necessary in language teaching. The problem is that, as Polguère (2004) has observed, right now, teachers are often times not trained enough to perform analysis of lexical phenomena quickly and accurately during class interactions. Lexicon is so large that there can’t be a ready-to-use answer to every possible question teachers can be asked. Our objective is therefore to make teachers more autonomous by training them to observe and to analyze lexical phenomena so they can face any new lexical problem. We believe that the knowledge of the participants to our experiments, all experienced lexicographers, could benefit the teachers as much as other lexicographers. For the time being, when analyzing the data from experiments, we describe participant’s every moves, without judging the efficiency or the relevance of each operation. Eventually, we’ll have gathered enough data to derive models of approaches to solve different types of problems, or to accomplish different types of tasks. Teachers could study these models, and pedagogical activities could be created for them. In so doing, not only would they learn by example, they could also initiate themselves to lexicography. Of course, the goal here is not to turn teachers into lexicographers, but to give them some basic tools and a method to analyze lexical phenomena efficiently. Another implication of our work for pedagogy could aim at learners; some activities could be derived from the ontology to train them using dictionnaires, help them understand the structure of definitions, the concept of synonymy, etc.

Acknowledgement

We’d like to thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for supporting this research, and the Département de linguistique et de traduction de l’Université de Montréal for a redaction bursary.

References

- Atkins, Sue & Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.
- Béjoint, Henri. 2000. *Modern Lexicography: an Introduction*. Oxford University Press. Oxford.
- Branch, Jennifer L. 2000. Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters. *Library & Information Science Research*, 22 (4): 371-392.
- De Pietro, Jean-François. 2003. L’enseignement du lexique en Suisse ou: comment en finir avec les listes à mémoriser? *Lettre de l’AIRDF*, 33: 12-18.
- Ericsson, Anders K. & Herbert A. Simon. 1984. *Protocol Analysis: Verbal Reports As Data*. MIT Press. Cambridge. MA.
- Ericsson, Anders K. & Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports As Data*. MIT Press. Cambridge. MA.
- Fontenelle, Thierry (editor). 2008. *Practical Lexicography: A Reader*. Oxford University Press. Oxford.
- Grossmann, F., Paveau, M.-A., Petit, G. 2005. *Didactique du lexique: langue, cognitions, discours*. ELLUG. Grenoble.
- Landau, Sidney. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press. Cambridge.
- Mel’čuk, Igor. 2006. Explanatory Combinatorial Dictionary. In G. Sica, editor. *Open Problems in Linguistics and Lexicography*: 225–355. Polimetrica. Monza.
- Mel’čuk, Igor, André Clas, & Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. Duculot. Paris/Louvain-la-Neuve. 1995.

- Noy, Natalya F. & Deborah L. McGuinness. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05* and *Stanford Medical Informatics Technical Report SMI-2001-0880*. Stanford University. Stanford. CA.
- Polguère, Alain. 1994. Savoir consulter un dictionnaire, c'est bien...Savoir comment on le construit, c'est mieux. *Québec français*, 134: 68-70.
- Russo, J. Edwards, Johnson, Eric J., & Stephens, Debra L. 1989. The Validity of Verbal Protocol. *Memory & Cognition*, 17 (6): 759-769.
- Simard, Claude. 1994. Pour un enseignement plus systématique du lexique. *Québec français*, 92: 65-67.
- Zgusta. Ladislav. 1971. *Manual of Lexicography*. Academia. Prague.