Investigating Semi-Automatic Procedures in Pattern-Based Lexicography

Laura Giacomini[1,2], Paolo DiMuccio-Failla

[1] Institute for Translation and Interpreting (IÜD), University of Heidelberg,
Plöck 57a, D-69117 Heidelberg

[2] Institute for Information Science and Natural Language Processing (IwiSt), University of Hildesheim, Universitätsplatz 1, D-31141 Hildesheim

E-mail: laura.giacomini@iued.uni-heidelberg.de, paolodimuccio@gmail.com

Abstract

In this contribution we present existing pattern description models with different degrees of computerization, discuss their potential from the perspective of the creation of an e-lexicographic resource for language learners, introduce the parameters of pattern accuracy and ontology reliability for a qualitative evaluation of the results, and make some proposals for a future quantitative evaluations. The models discussed are a) Hanks's CPA and the Pattern Dictionary of English Verbs (PDEV), b) methods employed by Tecling (Technologies for Linguistic Analysis, Pontifical Catholic University of Valparaiso, Chile) and Verbario, a pattern database of Spanish verbs, and c) an ongoing lexicographic project for the compilation of a learner's dictionary of Italian linked to a conceptual ontology. These approaches are founded in the tradition of theories focussing on the connection between lexis and grammar, especially in John Sinclair's view of *normal patterns of usage* as the true bearers of meaning of a language.

Keywords: pattern-based lexicography; semi-automatic procedures; ontology; pattern of usage; learner's dictionary

1. Introduction

Linguistic approaches covering, to different degrees, the interplay between lexical patterns and grammatical frameworks, or, in John Sinclair's words, "the meeting of lexis and grammar" (Sinclair, 1991: 81), have a quite long tradition ranging from lexicogrammar theories (cf. Halliday, 1992), to Gross's *classes d'objets* (1994) and Herbst's notion of *Konstruktikon* (2016). This tradition is largely intertwined with corpus-based and corpus-driven methods. In the context of pattern-based lexicography, especially in the sense of Sinclair (1991) and Hanks's Theory of Norms and Exploitations (Hanks, 2013), much research has been done to integrate notions of lexical semantics into the study of (phraseological) word combinations, giving birth to pioneering dictionaries such as the Collins COBUILD English Language Dictionary (COBUILD 1987) and the New Oxford Dictionary of English (Hanks & Pearsall, 1998).

However, methods for the computerization of the lexicographic process have been only recently taken into consideration as an essential part of pattern-centred dictionary research. In this contribution, we would like to compare existing semi-automatic pattern description models, discuss their potential from the perspective of the creation of an e-lexicographic resource for language learners, and make some proposals for improving work in the future.

This study belongs to the initial phase of our lexicographic project for the compilation of a learner's dictionary of Italian, for which the description of syntactic and semantic patterns of language has been chosen as the core microstructural criterion (DiMuccio-Failla & Giacomini, 2017a, 2017b). In the following, we will refer to the project as the IFL (Italian as a foreign language) project.

In the next section we first introduce the three models we are comparing in our study (Section 2). We then move to the relevant steps in the lexicographic process and corresponding solutions offered by the three models (Section 3). Finally, we discuss the impact of semi-automatic procedures on the lexicographic workflow and propose parameters for qualitative evaluation (Section 4).

2. Models

In this study we take into consideration three models for pattern-based lexicographic description. All these approaches originate from Sinclair's notion of *normal patterns of usage* as the true lexical units of a language: according to Sinclair, in general each major normal sense of a word can be associated with a distinctive pattern of usage determined by collocation, colligation, semantic preference and semantic prosody (e.g. Sinclair, 1996, 2004).

- Hanks' CPA and the Pattern Dictionary of English Verbs (PDEV) as its lexicographic result,
- Methods employed by Tecling (Technologies for Linguistic Analysis, a group of research in computational linguistics and NLP affiliated to the Pontifical Catholic University of Valparaiso, Chile) to automatically induce a taxonomy of nouns and generate patterns from corpora, and
- Experiments carried out within our lexicographic project for learners of Italian.

The Pattern Dictionary of English Verbs (PDEV) is the practical result of the application of Hanks' Theory of Norms and Exploitations (Hanks, 2013) and the technique of Corpus Pattern Analysis (CPA, Hanks, 2004b). The Pattern Dictionary of English Verbs is primarily intended as a resource for use in computational linguistics, due to its pattern formalization, but also in language teaching and cognitive science. It presently includes 1,423 complete verbs out of a total of 5,392. For each verb, a set of patterns is provided in which semantic types or semantic roles are indicated for each argument. Arguments in a pattern are linked to nodes in the CPA Ontology, a shallow semantic ontology which contains 253 semantic types.

Researchers at Tecling have taken Hanks' theory and the CPA's approach as a starting point to develop methods to automatically induce taxonomies of nouns and patterns of verbs from corpora. The language of application is Spanish. In the framework of the

Verbario project (2014-2017, www.verbario.com), a database of Spanish verbs was semi-automatically created. Verbario currently features two versions, one with manually created patterns, another with automatically generated patterns.

The IFL project is presently carried out by a group of researchers at Heidelberg University (Germany), Hildesheim University (Germany) and the University of Modena and Reggio-Emilia (Italy). We aim, on the one hand, at describing patterns of verbs and other word classes in a dictionary for learners of Italian and, on the other hand, at developing an ontology-like conceptual network in which semantic fillers (semantic types and roles) are collected, and on which lexicographic pattern description can be based. Our model has a clear cognitive orientation, in that it attempts to define word meanings by first identifying prototypical concepts and then finding and logically arranging related concepts. In the current, initial stage of the project, we are mainly concerned with studying patterns of different word classes, especially working on semantically homogeneous verbs. We also make some experiments in other languages (English, German, French), to test the validity of our method (cf. Orlandi, Giacomini & DiMuccio-Failla, 2019) and refine the results obtained for Italian.

3. Pattern-based lexicographic process and semi-automatic procedures

The models we intend to compare share, on the one hand, a common theoretical background, which has found application in different languages. On the other hand, they develop different strategies for the implementation of the core steps within the pattern-based lexicographic process: (a) detecting patterns in corpora, (b) selecting semantic types, (c) formally or informally expressing patterns, and (d) building taxonomies/ ontologies for semantic types. Moreover, they organize these procedural steps in different ways and choose different principles for sorting the meanings of polysemous words.

In this section, we describe and compare all the different strategies, especially from the perspective of computerization. For the purpose of this contribution, we will concentrate on verbal patterns only, since this is the main focus of the three models.

3.1 Identification of patterns and semantic fillers

PDEV, Verbario and the IFL dictionary all record data from corpora. For the PDEV, the British National Corpus has been used as the main reference corpus. Different to the PDEV, web corpora (esTenTen and itTenTen) have been used in Verbario and the IFL project for Spanish and Italian. Web corpora have the advantage of being large and heterogeneous enough to offer a broad spectrum of contexts, covering many different text genres and text types. On the negative side, at least for what concerns itTenTen in the focal project, the relatively great amount of noise and the imbalance in the distribution of text sources posed some problems. For these reasons, the IFL project also integrates in the lexicographic process a comparison of corpus data with existing general language and collocation dictionary data.

Table 1 summarizes the steps that enable the assignment of concordance lines to patterns:

CPA	**Tecling**	**IFL project**
Concordance sampling	Concordance sampling	Collocation extraction and concordance sampling
	Syntactic structures extraction	
	Semantic information extraction	
Sample analysis and pattern identification	Sample analysis and pattern identification	Sample analysis and pattern identification

Table 1: Process of pattern identification in the three models (blue field: manual step, grey field: semi-automatic step, white field: automatic step)

In the three models, concordance analysis delivers syntactic and semantic information about a verb in its contexts. However, in the IFL project collocation analysis is the starting point of investigation on which the analysis of concordances is based.

From the beginnings of the Sinclairian tradition in pattern-based lexicography, concordances have played a crucial role. An important issue concerns the appropriate number of concordance lines to be taken into consideration. Sinclair makes the case for *small samples*, a “screenful” of around 25 lines, which should be enough to get a first overview of the patterning of a node (2003: xiii-xiv). Analysis then continues in two possible directions: the main patterns can be confirmed by subsequently adding new small samples from the same dataset until no new information is obtained, or data can be refined if the initial search results are not satisfactory.

Hanks suggests that detailed analysis requires the selection of a random sample of up to 1,000 concordance lines, usually starting with a small sample of 200-250 lines (2004a: 255, PDEV). Tecling uses subsequent samples of around 100 corpus lines each, and note that a maximum of three samples is usually sufficient to identify all major patterns of a verb. Concordances are automatically generated, whereas the association of concordances with patterns is a manual step, usually carried out in an iterative way (see Figure 1 for the usual procedure). Hanks points out that “the identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art” (from the PDEV website).

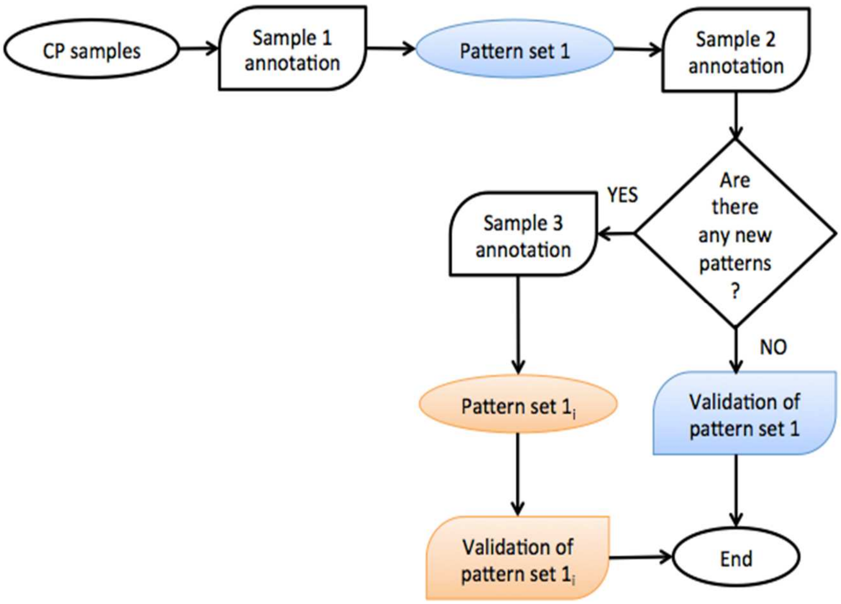
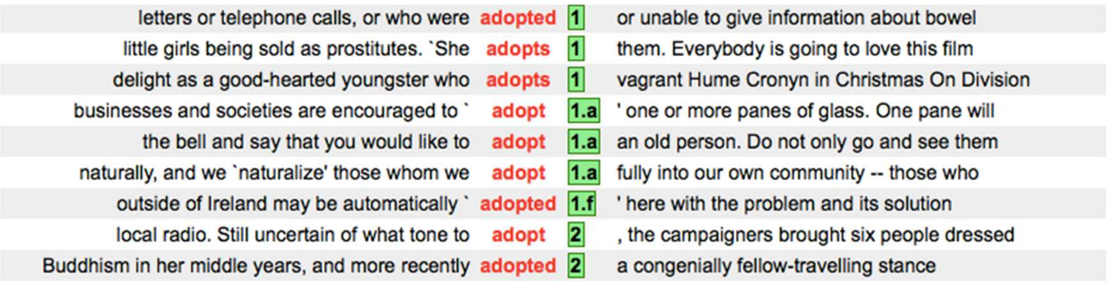


Figure 1: Iterative analysis of corpus concordances for pattern identification

CPA:

In CPA, one starts with concordance lines extracted by the SketchEngine concordancer (Kilgarriff et al., 2004), and groups them into semantic homogeneous sets, whereas “associating a ‘meaning’ with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns” (Hanks, 2004b: 88).

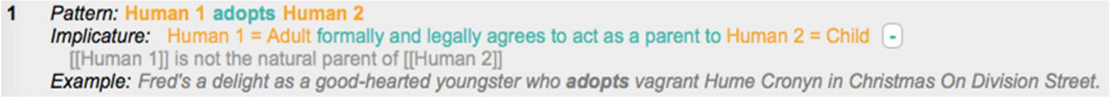
Each concordance line is manually annotated with a pattern number, exploitations and non-relevant data (e.g. errors or quotations) (Figure 2).



letters or telephone calls, or who were	adopted	1	or unable to give information about bowel
little girls being sold as prostitutes. `She	adopts	1	them. Everybody is going to love this film
delight as a good-hearted youngster who	adopts	1	vagrant Hume Cronyn in Christmas On Division
businesses and societies are encouraged to `	adopt	1.a	' one or more panes of glass. One pane will
the bell and say that you would like to	adopt	1.a	an old person. Do not only go and see them
naturally, and we `naturalize' those whom we	adopt	1.a	fully into our own community -- those who
outside of Ireland may be automatically `	adopted	1.f	' here with the problem and its solution
local radio. Still uncertain of what tone to	adopt	2	, the campaigners brought six people dressed
Buddhism in her middle years, and more recently	adopted	2	a congenially fellow-travelling stance

Figure 2: Annotation of concordance lines according to CPA

Once patterns have been identified, semantic values (types and roles) are manually attributed to the arguments of the input word in each pattern by referring to the CPA Ontology (cf. Section 3.3). One of the main issues of this step concerns the choice of the appropriate semantic values: “among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished” (from the PDEV website). As already mentioned in Section 2, each entry in the PDEV consists of a formalized pattern with its semantic fillers, an *implicature* expressing the pattern meaning in natural language (Hanks, 2004b: 88), a usage example and frequency indication (cf. Figure 3 for the core microstructural items).



1 *Pattern:* Human 1 adopts Human 2

Implicature: Human 1 = Adult formally and legally agrees to act as a parent to Human 2 = Child -

[[Human 1]] is not the natural parent of [[Human 2]]

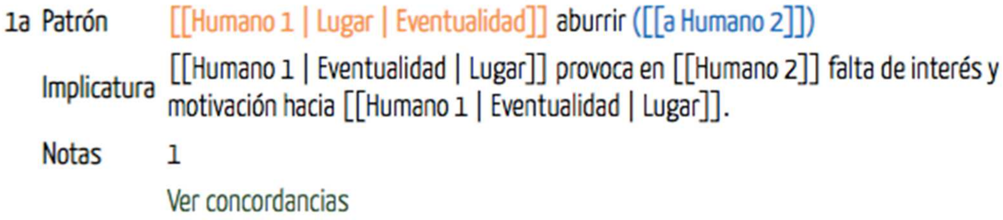
Example: *Fred's a delight as a good-hearted youngster who* ***adopts*** *vagrant Hume Cronyn in Christmas On Division Street.*

Figure 3: Entry example in PDEV

Tecling:

Concordances of a verb are extracted from the esTenTen corpus by means of Jaguar, a tool for corpus exploitation (http://www.tecling.com/jaguar). Concordance analysis is complemented with dependency analysis carried out by using Syntaxnet, Google's open-source parser. Semantic analysis also plays a role at this stage: named entities are classified through POL, a NER-tool for detecting and classifying names of geographical places, persons and organizations (http://www.tecling.com/pol), while common names are classified through a previously generated taxonomy (cf. Section 3.3). Patterns are identified on the basis of syntactic functions and semantic types. Experiments have been carried out to compare manual and automated pattern identification with the aim of improving automation in order to support lexicographers' work (Renau, et al., 2019; Renau & Nazar, 2016). Manual analysis of a set of verbs has been used as a gold standard to test the results of automatic analysis, in which semantic fillers are obtained from the available taxonomy. The main problem with the automatic output overspecification of semantic values for the arguments is that the implemented algorithm selects the first available semantic type by proceeding bottom-up in the taxonomy, frequently producing too specific and too many patterns (ibid.: 895-897).

The CPA orientation of this work is reflected by the entry structure in Verbario, for instance for the Spanish verb *aburrir* (*to bore*) (Figure 4):



1a Patrón [[Humano 1 | Lugar | Eventualidad]] aburrir ([[a Humano 2]])

Implicatura [[Humano 1 | Eventualidad | Lugar]] provoca en [[Humano 2]] falta de interés y motivación hacia [[Humano 1 | Eventualidad | Lugar]].

Notas 1

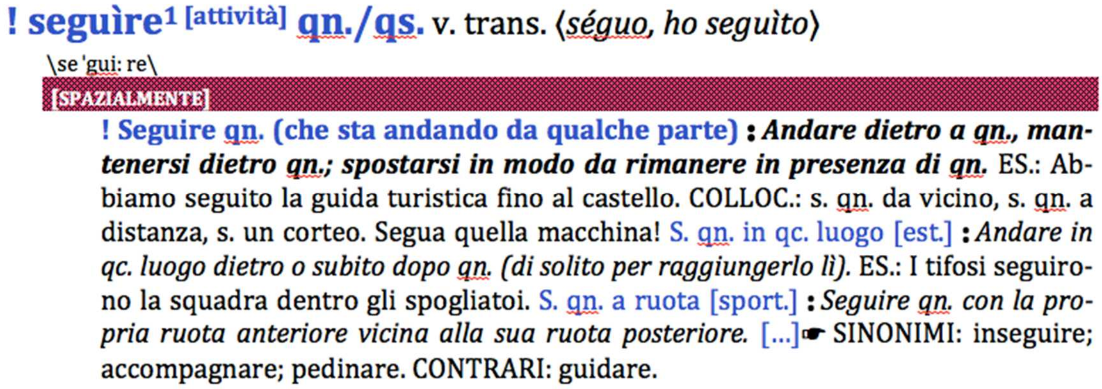
Ver concordancias

Figure 4: Entry example in Verbario

IFL project:

In the IFL project, we collect collocations of a node verb from the itTenTen corpus through the Sketch Engine word sketch tool, and then extract concordance lines referring to these collocations in order to validate them. Random corpus samples filtered by using the GDEX function (Kilgarriff et al., 2008) are analysed according to the meaning of the node verb and patterns are thus gradually identified.

Lexicographic work in the IFL project has a clear phraseological orientation: not only do we firmly believe that patterns are phraseological in nature (Sinclair, 1991, 1996), but we also explore collocations to identify, validate and refine our patterns and, in general, we use phraseological disambiguators to cluster patterns with close meanings (DiMuccio-Failla & Giacomini, 2017b). Collocation analysis sheds light on the syntactic, semantic, and phraseological features of verbs at the same time. Figure 5 shows the informal pattern description in a possible data presentation mode of the planned IFL dictionary.



! seguìre^{1} [attività] qn./qs. v. trans. ⟨*séguo, ho seguìto*⟩

\se 'gui: re\

[SPAZIALMENTE]

! Seguire qn. (che sta andando da qualche parte) : *Andare dietro a qn., mantenersi dietro qn.; spostarsi in modo da rimanere in presenza di qn.* ES.: Abbiamo seguito la guida turistica fino al castello. COLLOC.: s. qn. da vicino, s. qn. a distanza, s. un corteo. Segua quella macchina! S. qn. in qc. luogo [est.] **:** *Andare in qc. luogo dietro o subito dopo qn. (di solito per raggiungerlo lì).* ES.: I tifosi seguirono la squadra dentro gli spogliatoi. S. qn. a ruota [sport.] **:** *Seguire qn. con la propria ruota anteriore vicina alla sua ruota posteriore.* [...]☛ SINONIMI: inseguire; accompagnare; pedinare. CONTRARI: guidare.

Figure 5: Entry example (*seguire, to follow*) in the planned IFL dictionary

Challenges typically encountered at this stage are:

- In the case of the IFL project, first grouping collocations into semantically homogeneous sets, each identifying a pattern.
- Distinguishing primary patterns from secondary patterns.
- Assigning semantic values to argument slots: selection restrictions and determining the appropriate degree of generalization.

3.2 Pattern sorting

Interestingly, the sorting of patterns in the final application of the three models (PDEV, Verbario, IFL project) complies with different principles. In the PDEV, the patterns of a verb are sorted according to their cognitive salience. Also in the IFL project, senses identified by patterns are sorted according to their cognitive relevance: we start from the idea of a conceptual network in which the related senses of a word are organized in a radial set around one or possibly more prototypical concepts. This assumption verifies the cognitivist account of polysemy proposed by Brugmann and Lakoff (cf., among others, Brugmann & Lakoff, 1988). The fundamental meaning of a verb is thus followed by other senses linked by metonymy, abstraction, and metaphor relations (cf. examples in DiMuccio-Failla & Giacomini, 2017b).

In the Tecling project, patterns are sorted by decreasing order of frequency (Renau & Nazar, 2016: 827). This sense enumeration approach has been criticized in the past, not least in the context of the COBUILD Dictionary, because it often leads to unnatural results (cf. Lew, 2013; DiMuccio-Failla & Giacomini, 2017b). We think that the cognitive criterion of sense disambiguation is the most suitable way of presenting meanings of polysemous words to language learners, since it logically guides the dictionary users from a prototypical meaning towards all related senses (e.g. figurative senses).

3.3 Ontology building

The role of an ontology of semantic types and semantic roles in pattern-based lexicography is of crucial importance: the systematic conceptual classification of these items guarantees consistency in their use throughout the dictionary and potentially simplifies pattern formulation. In this contribution, for reasons of simplicity, we will use the term *ontology* to refer to a typically hierarchical structure of entities or concepts, irrespective of its complexity and degree of expressiveness, therefore also including taxonomies. The three discussed models show clear differences with regard to

- the method for ontology building and
- the way in which the ontology interfaces with pattern identification.

Table 2 shows the role of the ontology within the process of pattern identification in the three models:

CPA	**Tecling**	**IFL project**
CPA Ontology	Taxonomy	
Concordance sampling	Concordance sampling	Collocation extraction and concordance sampling
	Syntactic structures extraction	
	Semantic information extraction	
Sample analysis and pattern identification	Sample analysis and pattern identification	Sample analysis and pattern identification
		Conceptual ontology

Table 2: Ontology and pattern identification in the three models (blue field: manual step, grey field: semi-automatic step, white field: automatic step)

CPA:

The CPA Ontology is based on work done by Pustejovsky et al. (2004). It is a shallow semantic ontology created by progressively compiling and organizing a list of semantic types (El Maarouf, 2013). As previously indicated, in the PDEV each argument of each pattern is linked to a node in the CPA Ontology, which can be accessed via the dictionary website. Here is a brief example of a hierarchy:

Anything > Entity > Physical Object > Inanimate > Artefact > Building

Final nodes of the ontology may be very specific, but sibling concepts are still missing. For instance, the only two available subcategories of Building are Cinema and Theatre, whereas for Food only the subcategoy Meat is given.

Tecling:

Tecling uses a statistically-based taxonomy induction algorithm to generate a taxonomy of Spanish nouns from a corpus. Different quantitative approaches are simultaneously applied, among which the computation of similarity coefficients to identify sibling words and of asymmetric co-occurrence to find parent-child nodes (Nazar & Renau, 2016). As pointed out in Renau and Nazar (2016), this procedure relies on an existing taxonomy structure. In fact, semantic types contained in the CPA Ontology provide the conceptual architecture into which around 35,000 Spanish nouns are automatically inserted. The results are compared with the Spanish WordNet 1.6 (Atserias et al., 2004), which serves as a gold standard (for a brief discussion on the use of wordnets as sources for semantic types, see further down in this section). Insights into the automatically induced taxonomy are provided by the ontology webpage (www.tecling.com/kind). For instance, if we search for the category Comida (Food), we get a full list of four hypernyms and 157 hyponyms. The taxonomy induction algorithm employed by Tecling can detect both symmetric and asymmetric relations, and achieves an estimated average of 77.86% precision and 33.72% recall on the total results (Nazar & Renau, 2016).

IFL project:

In our project, a conceptual ontology is developed alongside the process of pattern identification. It is important to note that we presently work on patterns without employing an external ontology. Instead, we build a new conceptual network according to a bottom-up procedure, in which semantic types (and lexicalized semantic roles) selected for patterns are progressively fed into the ontology.

For instance, one of the patterns of the verb *seguire* (*to follow*) is

seguire il racconto, la spiegazione o l'argomentazione di qn.

We insert the semantic types *Racconto* (Narration), *Spiegazione* (Explanation) and *Argomentazione* (Argumentation) into our ontology and link them to other concepts, e.g. synonyms such as *Narrazione* (Narration) and hypernyms, in this case via a polyhierarchical structure, in the sense that the three semantic types have two different hypernyms, *Evento comunicativo* (Communicative event) and *Rappresentazione formale di un evento comunicativo* (Formal representation of a communicative event)

(Figure 6):

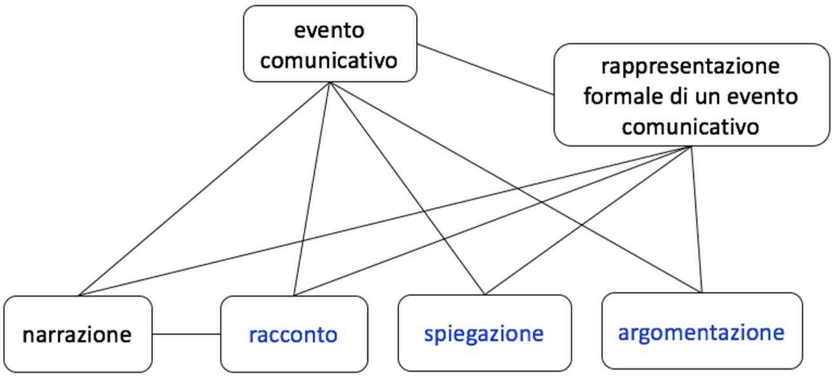


Figure 6: Excerpt from the ontology, with semantic types derived from pattern formulation

In order to systematically detect relevant types, we analyse clusters of semantically close verbs (e.g. synonyms, converses, or troponyms), which display some meaning overlap and are likely to share a number of semantic fillers. As we are still at an initial stage of the project, we only have a very small number of items in our ontology, corresponding to a small number of words in the dictionary. As the ontological structure is being configured, its items are used in a top-down procedure to fill argument slots of new verbs. Being dependent on their usage as semantic fillers in argument slots, the hierarchy of types only has to be as systematic and coherent as normal language usage.

The ontology is not only a repository of semantic types, it also provides a clear overview of the lexical domains we intend to cover and facilitates consistent dictionary definitions. The upper part of the ontology draws on the EuroWordNet model (Vossen et al., 1998), which, in turn, is based on Lyon's (1977) tripartite entity categorization. In the lower part of the ontology entities are further classified into types (cf. DiMuccio-Failla & Giacomini, 2017a). Tests performed on ItalWordNet and experiments carried out on English, German and French using the Princeton WordNet, GermaNet and WoNeF, reveal that wordnets have a limited reliability with regard to semantic types: they pose major problems for meaning disambiguation (for instance, synsets are not always clearly distinct from each other). Moreover, they often introduce scientifically motivated subcategorizations "that are not in ordinary usage" (Jezek & Hanks, 2010) and therefore not useful for lexicographic purposes[1] (wordnets' drawbacks in this sense have also been described by Hanks and Pustejovsky (2005) and Renau et al. (2019)). As pointed out by Polguere, the Princeton WordNet's ontological structure "is not as cognitively relevant as it was expected to be by its designers [...], [since] the focus of the project shifted at an early stage from psycholinguistics to computer applications"

[1] Jezek & Hanks (2010) also see a problem in the attempt to force all items of a language into a taxonomic hierarchy.

(2014: 397). This aspect, which appears to be common to all wordnets, is crucial to our approach to lexicography, which, instead, has a strong cognitive orientation.

The challenge typically encountered at this stage is:

- No conceptual ontology is already available from which semantic types can be reliably obtained.

4. Semi-automatic procedures and lexicographic workflow: qualitative analysis

We will now concentrate on the impact of automatic procedures on time efficiency and the quality of the lexicographic results, for instance on the accuracy of patterns and reliability of the underlying ontology. In the previous sections we introduced the set of automatic steps used either in all three models or only in some of them:

- taxonomy induction (Tecling)
- concordance sampling (CPA, Tecling, IFL project)
- syntactic structures extraction (Tecling)
- semantic information extraction (Tecling)
- collocation extraction (Tecling, IFL project)
- pattern extraction (IFL project)

We attempt to assess the potential of these methods specifically for the production of a learner's dictionary, which is the main goal of the planned IFL project but not of the two other models. The results of CPA and Tecling research, namely PDEV and Verbario, are in fact rather to be understood as databases in which formal data representation can serve as a possible source for a learner's lexicography.

Due to the differences between the described models (e.g. language, degree of computerization, intended goal), at the moment we cannot rely on any metrics for a quantitative evaluation of the results. Even within the same model, a quantitative evaluation is a difficult goal to achieve. As pointed out by Renau et al. (2019: 897) in the case of automatic pattern generation, for example, it is even impossible to establish a baseline, since we are not dealing with a classification system in which a certain chance of success with a random selection or a trivial method is given.

We therefore provide a primarily qualitative analysis based on the examination of the achieved results (pattern accuracy, also in comparison to monolingual dictionaries, and ontology reliability), and observations made by the involved researchers about their own work. The parameters we chose for assessing the quality of the final results are pattern accuracy and ontology reliability. Details regarding final results according to these parameters will now be presented, followed by remarks on time efficiency and source data.

4.1 Pattern accuracy and ontology reliability

	PDEV	Verbario	IFL project
Sample	*follow, need, choose, adopt, eat*	*abrir, aburrir, acentuar, activar, cortar*	*seguire, inseguire, accompagnare, pedinare, incalzare*
Pattern uniqueness	Patterns are distinct from each other	Patterns are not always distinct from each other	Patterns are distinct from each other
Pattern expressiveness	Heterogeneous degree of expressiveness: several semantic fillers appear to be too generic	Generally limited degree of expressiveness	High degree of expressiveness: semantic fillers are as specific as possible
Semantic coverage	Large semantic coverage, almost all dictionary senses corresponding to normal usage match a pattern (Dictionaries: COBUILD, ODE)	Large semantic coverage, almost all dictionary senses corresponding to normal usage match a pattern (Dictionaries: DAELE, SALAMANCA)	Each dictionary sense corresponding to normal usage matches a pattern (Dictionaries: TRECCANI, DE MAURO)
Ontology depth	Shallow ontology with limited inheritance levels	This kind of taxonomy appears to have a greater depth than the CPA Ontology	The depth of the ontology depends on normal language usage (bottom-up approach)
Relation patterns-ontology	Top-down approach: coherent usage of semantic types in patterns according to the depth of the ontology	Top-down approach: coherent usage of semantic types in patterns according to the depth of the ontology	Bottom-up approach: the ontology is systematically filled with semantic types selected during pattern identification

Table 3: Pattern accuracy and ontology reliability in the three models

Pattern accuracy is tested by selecting a small verb sample from each dataset and considering, for each verb, the uniqueness of patterns (each pattern identifies one distinct meaning), the expressiveness of semantic types used as argument slot fillers (degree of generalization), and semantic coverage in comparison to meaning presentation in existing monolingual learner's dictionaries[2]. Ontology reliability is tested by the conceptual depth of the ontology and the way in which the ontology interfaces with the building of dictionary patterns. Table 3 shows the results of our analysis. Verbario has been considered in its manual version, since the automatically generated verb entries cannot be presently accessed online.

The clearest difference between the lexicographic results obtained by the three models concerns the expressiveness of patterns, and the depth of the ontology (for the important factors here see the examples mentioned in Section 3.3). Pattern expressiveness seems to be more dependent on the chosen approach rather than on process automation, which explains the similarity between data in the PDEV and Verbario as opposed to the IFL project data. These observations hint at the fact that the correlation between computerization, on the one hand, and pattern accuracy and ontology reliability on the other, should not be overrated in any direction.

4.2 Time efficiency and initial data

Some remarks need now to be made on time efficiency: generally speaking, time efficiency is enhanced by the application of any automated procedures. However, the balance between the amount of time saved thanks to automatic data extraction and the amount of time spent to correct and prepare data for presentation in a dictionary should also be taken into account (cf. also Renau et al. (2019) on the comparison between manual extraction and automatic extraction of patterns).

In our experience, manual work for pattern identification and ontology building requires a considerable amount of time, but this process can be significantly accelerated as soon as targeted initial data are available, for instance complex collocations, or semantic and pragmatic information found in discourse (e.g. stage-dependent conditions for a verb's meanings, cf. Kratzer (1995)). We are presently investigating methods for creating automatic procedures that are able to provide this kind of raw data. In addition to this, cross-language experiments with English, German and French help us refine both ontological and lexicographic data (a multilingual approach has been partly adopted in the context of CPA as well, cf. Baisa et al. (2016)). The manually compiled conceptual ontology in the IFL project may serve in the future as a gold standard for a quantitative evaluation of automatically obtained results, not only for Italian but also for other languages.

[2] There is no learner's dictionary for Italian yet, thus we had to use general monolingual dictionaries.

5. Conclusions

The idea of a direct comparison with similar methods originated in issues encountered during our empirical work on patterns. These issues can be summarized as follows:

- Detecting patterns in corpora and validating them against the content of existing dictionaries requires a considerable amount of time, especially for extracting relevant data such as syntactic structures and collocations.

- Formulating patterns (either informally or formally) is a conceptually complex activity; especially the choice of adequate semantic types and roles for argument slots would be easier if a corresponding ontology was already available.

- Building such an ontology is closely related to the building of patterns. Due to the impossibility of using existing ontologies or wordnets as a source for cognitive conceptual information (cf. Section 3.3), this task is also particularly challenging, especially for what concerns the selection of the appropriate generalization level.

All these issues greatly affect the lexicographic workflow in the IFL project, and will presumably remain at the heart of the discussion also in the future. We assume that automation can only improve the workflow in a satisfactory way as long as it does not require much manual effort to correct data at a later stage. As shown in the previous sections, much depends on the theoretical approach to pattern description. For the moment, the degree of pattern expressiveness and cognitive consistency aimed at in the IFL project can only be achieved by native speaker introspection (on the importance of introspection and intuition, cf. Sinclair (1991, 2004)). Introspection, however, benefits from the availability of suitable, automatically extracted initial data and will be further enhanced as soon as the linked conceptual ontology reaches a sufficient level of completeness to be used to automatically detect patterns of similar verbs in corpora.

6. References

Atserias, J., Villarejo, L. & Rigau, G. (2004). Spanish WordNet 1.6: Porting the Spanish WordNet across Princeton versions. In N. Calzolari, K. Choukri & T. Lino et al. (eds.) *Proceedings of the Fourth International Conference on Language and Resources Evaluation* (LREC), pp. 161-164.

Baisa, V., Može, S., & Renau, I. (2016). Multilingual CPA: Linking Verb Patterns across Languages. In T. Margalitazde & G. Meladze (eds.) *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 410-417.

Brugman, C. & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small*, G. Cottrell & M. Tanenhaus (eds.)* Lexical Ambiguity Resolution. San Mateo, California: Morgan Kaufmann, pp. 477–507.

COBUILD (1987): *Collins COBUILD English Language Dictionary.* Collins.

COBUILD: *Collins COBUILD Advanced Learner's Dictionary (2014).* Harper Collins.

DAELE. *Diccionario de aprendizaje del Español como Lengua Extranjera.* Accessed at: http:// http://www.iula.upf.edu/rec/daele (10 June 2019)

DE MAURO (2019). Il Nuovo De Mauro. Accessed at: https://dizionario.internazionale.it (10 June 2019)

DiMuccio-Failla, P. V. & Giacomini, L. (2017a). Designing an Italian learner's dictionary based on Sinclair's lexical units and Hanks's corpus pattern analysis. In I. Kosem et al. (eds.) *Proceedings of the Fifth eLex Conference Electronic Lexicography in the 21st Century.* Leiden, Netherlands.

DiMuccio-Failla, P. V. & Giacomini, L. (2017b). In M. Mitkov (ed.) *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017*, LNAI 10596. Springer, pp. 290-305.

El Maarouf, I. (2013). Methodological Aspects in Corpus Pattern Analysis. *ICAME Journal*, 37, pp. 119-148.

Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, pp. 15-30.

Halliday, M. A. K. (1992). Some lexicogrammatical features of the zero population growth text. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text.* Amsterdam: Benjamins, pp. 327-358.

Hanks, P. (2004a). The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3), pp. 245-274.

Hanks, P. (2004b). Corpus pattern analysis. In *Proceedings of the XI EURALEX International Congress*, Vol. 1, pp. 87-98.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations.* MIT Press.

Hanks, P. & Pearsall, J. (eds.) (1998). *New Oxford Dictionary of English*, 1st ed. Oxford University Press, Oxford

Hanks, P., & Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, *10*(2), pp. 63-82.

Herbst, T. (2016). Wörterbuch war gestern. Programm für ein unifiziertes Konstruktikon. In S. J. Schierholz, R. H. Gouws, Z. Hollós & W. Wolski (eds.) *Wörterbuchforschung und Lexikographie.* Berlin/Boston: de Gruyter.

Jezek, E., & Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, *4*(7).

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The sketch engine. *Information Technology*, pp. 105-116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, pp. 425-431.

Kratzer, A. (1995). Stage-level and individual-level predicates. In G. N. Carlson & F. J. Pelletier (eds.) *The generic book*, Chicago. University of Chicago Press, pp. 125-175.

Lew, R. (2013). Identifying, ordering and defining senses. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography.* Bloomsbury, pp. 284–302.

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

Nazar, R. & Renau, I. (2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. In N. Calzolari et al. (eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16).* European Language Resources Association (ELRA).

ODE 2019: *Oxford Dictionary of English.* Accessed at: https://en.oxforddictionaries.com. (05 June 2019)

Orlandi, A., Giacomini, L. & DiMuccio-Failla, P. V. (2019). I disambiguatori fraseologici nella lessicografia di apprendimento: una proposta per l'italiano e il francese. *Repères-Dorif*, 18.

PDEV. *Pattern Dictionary of English Verbs.* Accessed at: http://pdev.org.uk (10 June 2019)

Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, *27*(4), pp. 396-418.

Pustejovsky, J., Hanks, P., & Rumshisky, A. (2004). Automated Induction of Sense in Context. *COLING 2004.* Geneva, Switzerland.

Renau, I. & Nazar, R. (2016). Automatic Extraction of Lexical Patterns from Corpora. In T. Margalitazde & G. Meladze (eds.) *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 823-830.

Renau, I., Nazar, R., Castro, A., López, B., & Obreque, J. (2019). Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos. Estudios de Lingüística*, *52*(101).

SALAMANCA (2006) Cuadrado, J. G. *Diccionario Salamanca de la lengua española:[español para extranjeros].* Santillana.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.

Sinclair, J. (2003). *Reading concordances: An introduction.* Pearson Longman.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse.* Routledge.

TRECCANI (2019). Vocabolario Treccani. Accessed at: http://www.treccani.it/vocabolario (05 June 2019)

Verbario. Sémantica de los verbos en contexto. Accessed at: http://www.verbario.com (05 June 2019)

Vossen, P. et al. (1998). The EuroWordNet Base Concepts and Top Ontology. Document LE2-4003, D017, D034, D036, WP5. Accessed at: http://dare.ubvu.vu.nl/bitstream/handle/1871/11130/D017.pdf (20 March 2019)