Polysemy index for nouns: an experiment on Italian using the PAROLE SIMPLE CLIPS lexical database

Francesca Frontini*, Valeria Quochi*, Sebastian Padó, Jason Utt**, Monica Monachini***
*ILC CNR Pisa, **University of Stuttgart
name.surname@ilc.cnr.it, **[pado—uttjn]@ims.uni-stuttgart.de

Abstract

An experiment is presented to induce a set of polysemous basic type alternations (such as ANIMAL-FOOD, or BUILDING-INSTITUTION) by deriving them from the sense alternations found in an existing lexical resource. The paper builds on previous work and applies those results to the Italian lexicon PAROLE SIMPLE CLIPS. The new results show how the set of frequent type alternations that can be induced from the lexicon is partly different from the set of polysemy relations selected and explicitly applied by lexicographers when building it. The analysis of mismatches shows that frequent type alternations do not always correspond to prototypical polysemy relations, nevertheless the proposed methodology represents a useful tool offered to lexicographers to systematically check for possible gaps in their resource.

Keywords: polysemy, lexical resources, semantics

1. Introduction

The research presented in this paper deals with the problem of lexical ambiguity, building upon and extending previous work by Utt and Padó (2011). There a methodology was described for deriving systematic alternations of senses from a lexical resource in order to automatically classify words whose sense variation is predominantly polysemous or homonymous. The present paper exploits this work and sets itself two goals:

(a) Apply the methodology to a different language, namely to Italian instead of English, and to a different resource, the PAROLE SIMPLE CLIPS lexicon (PSC), Lenci et al 2000, instead of WordNet. Our specific interest in PSC is due to its being linked to a bespoke ontology, the SIMPLE ontology (Toral and Monachini, 2007), which is common to all lexica of the European SIMPLE project[1] and which contains a set of ontological types which can replace those used in the original experiment (Corelex, Buitelaar (1998));

(b) Test whether the inventory of regular polysemy relations already encoded in the PSC semantic layer can be extended using the model's predictions, and whether the proposed methodology can help identify gaps or inconsistencies in the resource [2].

First of all, a general outline of the problem of lexical ambiguity is given; then the first experiment of Utt and Padó, upon which the present paper builds, is summarized; later the PSC lexicon is described and its ontology is compared to the one defined by Corelex; finally the experiments are described and discussed.

2. The problem

Lexical ambiguity refers to the phenomenon of a word having more than one sense. Two main categories of lexical ambiguity are generally recognized: homonymy and polysemy.

2.1. Homonymy

The definition of homonymy slightly differs between theoretical linguistics and computational linguistics. The most common definition of homonymy in theoretical linguistics is that two words are homonymous if they share the same form (orthography and/or phonology), but have different, unrelated and mutually underived meanings (Leech, 1974; Lyons, 1977; Saeed, 1997). According to this view, two homonymous words must have different etymologies. Pure homonyms, moreover should manifest both homophony and homography.

In computational linguistics and in particular in lexicographic approaches that deal with written text, however, homonymy usually encompasses both strict homonymy and homography, given that when it comes to treating text there is no *a priori* means of distinguishing the two. Two words are homonymous according to this view when they share the same form, but have unrelated meanings.

Traditionally, lexicographers tend to identify etymological or historical unrelatedness as the main criterion for identifying homonyms. This etymological criterion is however highly questionable as a valid methodology for the synchronic description of languages, especially for computational approaches, since diachronic derivation is not an operational criterion and not necessarily part of a speaker's awareness (Zgusta, 1971).

2.2. Polysemy

The notion of polysemy has received ample treatment in the literature (Apresjan, 1974; Nunberg and Zaenen, 1992; Copestake and Briscoe, 1995; Nunberg, 1995; Palmer, 1981).

We will summarize various different approaches, simplifying things for our current purposes by only considering the following types of polysemy:

1. Regular (or logical polysemy): words with two, or

[1] SIMPLE was originally developed for 12 EU languages, and today some lexicons are also available in RDF formats.

[2] This latter point goes bejond the scope of Utt and Padó (2011).

more, systematically related meanings. The meaning of a word is described here in terms of the semantic (or ontological) classes to which the senses of a lexical item refer. Regular polysemy can thus be defined in terms of regularity of type alternations, where the "alternating types" in question are the semantic and ontological categories to which the senses of a lemma belong (Palmer, 1981; Pustejovsky, 1995).

So for instance, the fact that that a word used to describe an ANIMAL can also be used to describe its MEAT is referred to as the ANIMAL−FOOD alternation. This alternation is applicable to all (or most) instances of the semantic class, or type ANIMAL (e.g. *manzo, pollo, tacchino*, 'cow/beef', 'chicken', 'turkey', etc.); the regularity of the alternation is shown by its productivity, which extends to animals the use of which for meat might not be obvious in a given linguistic community (thus it is grammatical for an Italian to say *Ho mangiato il serpente in Cina e mi piaciuto* 'I ate snake in China and I liked it').[3].

These systematic meaning alternations are generally salient on conceptual grounds, common to (several) other words, usually derivable by metonymic sense shifts. For our current purposes, regular polysemy can thus be defined as the property of a word having distinct ordinary meanings referring to different objects, or ontological types, that bear some (metonymic) relation with one another.

2. Occasional (or irregular) polysemy: a word shows a "derivable" meaning alternation, i.e. there is an evident relation between the meanings, usually again metonymic, but this is not pervasive in the language (e.g. *coccodrillo*, 'crocodile', can be used both to indicate the animal and the (leather) material; this alternation is common to other animal words but is not so pervasive, and is clearly dependent on other world-knowledge factors)

3. Metaphorical polysemy: a word with meanings that are related by some kind of metaphorical extension. Again, this will not be systematic in the language, although other words may show similar extensions. For example, *fulmine*, 'lightning' NATURAL PHENOMENON, can be used metaphorically to described something or someone as 'very fast' as in *Giovanni è un fulmine*, 'John is as quick as a flash'; *Boa*, 'boa', ANIMAL, can also refer to a feather scarf. The relationship between the two senses of these words is probably one of lexicalized metaphorical extension which it will be hard to generalize to other words.

Polysemy is generally identified as a lexical-semantic phenomenon, clearly distinguished from similar "pragmatic" ambiguity cases. For instance, it is a pragmatic phenomenon, not a semantic one, that in many situations people can be identified by the name of an artifact that is temporarily associated with them, as in *red shoes is angry today* or *the ham sandwich would like to pay*, whereas the systematic use of the name of a container to refer to its content is identified as a semantic pheomenon (as in *I drank a glass with a friend yesterday*).

An appropriate treatment of polysemy (and homonymy) has clear lexicographic consequences: a lexical resource would be flawed if it failed to capture the fact that a lemma such as *rabbit* is subject to a certain type of systematic polysemy.

The distinction between regular polysemy, occasional polysemy and homonymy is somewhat more blurred than it seems at first (Zgusta, 1971; Palmer, 1981; Lyons, 1977; Landau, 1984; Ndlovu and Sayi, 2010), and a continuum can be recognized. On one end of this spectrum are those words whose senses are totally unrelated and which present very rare type alternations; then we find cases of unsystematic polysemy, with related senses, but which still present rare type alternations; finally there is systematic polysemy, with clearly related senses and a type alternation that systematically occurs in the lexicon of that language. As for unrelatedness, the traditionally established diachronic criterion is not undisputed, is not easy to apply even for linguists, is not necessarily salient synchronically for native speakers and is above all not easily operationizable. Returning to the example of *boa*, the metaphorical relationship between the 'scarf' and 'snake' senses may not be evident to all speakers, thus making it a case of homonymy instead of occasional polysemy. The frequency of a type alternation across lemmas in the lexicon on the other hand appears to be a clear and recognizable criterion for modelling the distinction between polysemy and homonymy[4].

3. Related Work

Utt and Padó (2011) start from the idea of (coarse-grained) ontological classes. If we consider a lemma and all of its senses, each possible sense can be labelled with an ontological class or type and thus each pair of senses of that lemma can be seen as an alternation between two ontological types. Such alternations are called basic alterations (BAs). An instance of BA (i.e. a sense pair within a lemma) may represent a case of regular (systematic) polysemy or a case of simple homonymy. However, when the same BA occurs across many lemmas, this can be taken as evidence of a regular polysemy.

For example, in languages such as English or Italian the presence of a large number of lemmas with two senses, one of which is labelled with the type ANIMAL and the other with the type FOOD provides evidence of the fact that the FOOD#ANIMAL BA is not merely sporadic in such languages but the product of ANIMAL >FOOD regular polysemy.

Utt and Padó (2011) use frequency to distinguish polysemous vs. homonymous BAs derived from WordNet senses

[3]Notice that languages often show lexical gaps, as with the English *cow/beef* "exception", so that a lexical-rule approach has to deal with these as such (Nunberg, 1995; Copestake and Briscoe, 1995).

[4]Admittedly, the frequency criterion is quite simplistic. In addition, treating the issue of homonymy/polysemy at the lemma level is an oversimplification as it fails to account for lemmas with more than two senses only one of which is unrelated to the others. However, it has several practical advantages for NLP

tagged with the 39 ontological types provided by CoreLex. In the following paragraph a precise outline of the algorithm will be given, but the basic intuition runs as follows. Assuming that a list of polysemic BAs were already available, the polysemy index of a lemma w can be calculated as the ratio of polysemic BAs over the total BAs of the lemma:

$$\pi_n(w) = \frac{P_N \cap P(w)}{P(w)} \quad (1)$$

where $P(w)$ is the set of BAs for that lemma and P_N is the set of BAs that are classified as regular polysemies. So if a lemma (*bank*) has three senses (typed as INSTITUTION, BUILDING, PART_OF_RIVER) and instantiates three BAs, one of which is listed as a regular polysemy (BUILDING>INSTITUTION) and two aren't, the polysemy index for that word will be 1/3. Following on from this assumption, the optimal frequency threshold N for a BA to be classified as a regular polysemy will be one which assigns a higher index to typically polysemous lemmas and a lower one to typically homonymous ones. Thus the authors use equation (1) and a set of 24 homonymous and 24 polysemous English nouns drawn from the literature to derive the list of N most frequent BAs which are to be considered regular polysemies.[5].

4. Using Parole Simple Clips

In our experiment, we intended to replicate the homonymy-polysemy distinction experiment for Italian, exploiting PAROLE SIMPLE CLIPS (PSC), a multi-layered lexicon (Ruimy et al., 1998) for Italian. The lexical information in PSC is encoded at different descriptive levels: at the phonetic, morphological, syntactic and semantic layers. The semantic layer of PSC, SIMPLE (Lenci et al., 2000) is largely based on Pustejovsky's Generative Lexicon theory (Pustejovsky, 1995).

In PSC, each lexical entry is organized into senses, called *usems*, that in are turn linked to each other by a rich set of semantic and lexical relations. Within the scope of this work the relations that are most interesting to us are those linking two usems belonging to the same lexical entry, among which, crucially for this experiment, are relations of regular polysemy. The regular polysemous classes represented in SIMPLE were elaborated by starting from a list proposed by Wim Peters and enriched through the main regular polysemies listed in Malmberg (1988).

Moreover each sense is defined by its membership in an ontological type. Such types are drawn from a crosslingual ontology[6], with labels in English, that was defined and used by all SIMPLE projects in different languages. The ontology is hierarchical and the types used in the lexicon represent the nodes of the ontology.

To sum up, the information in PSC enables us to:

1) identify cases of semantic ambiguity by listing lexical entries with more than one usem, such as:

maiale ("pig") and *boa*;

2) retrieve the ontological type of each usem, and list/count instances of Basic Alternations (BAs), e.g.:

- the pair (USem01934maiale as SUBSTANCE_FOOD, USem1933maiale as EARTH_ANIMAL) is an instance of the SUBSTANCE_FOOD#EARTH_ANIMAL BA
- the pair (USem4328boa as EARTH_ANIMAL, USem67540boa as CLOTHING) is an instance of the EARTH_ANIMAL#CLOTHING BA;

3) retrieve the encoded relations between usems; e.g.:

- the pair (USem01934maiale–USem1933maiale) is linked by an encoded relation of *PolysemyAnimal-Food*;
- the pair (USem4328boa–USem67540boa) has no encoded relation;

In terms of the aforementioned methodology, the bottom-up extraction of BAs can be performed by using the information about lexical entries, usems and their types contained in SIMPLE. On the other hand the explicitly encoded lexical-semantic relations between usems of the same word can be used for:

- the selection of the set of homonymous and polysemous lemmas that are required for the procedure of induction of the optimal frequency threshold for BAs;
- the evaluation of the induced threshold.

Apart from regular polysemy, two sets of explicit relations are relevant for this experiment. Firstly, metaphor, such as the relation linking *asino* as HUMAN (a stupid person) and *asino* as ANIMAL (a donkey); metaphoric extension is a less systematic phenomenon than regular polysemy, but the presence of such relations between two usems of a word can be viewed as evidence of the fact that the two senses are not unrelated. Secondly, qualia structure relations - Constitutive, Telic, Agentive, Formal (Pustejovsky, 1995); for instance *anice* as ARTIFACTUAL_DRINK *is-made-of anice* as FRUIT ('anise'). The qualia structure has often been defined as that part of general encyclopedic world knowledge that is accessible or relevant for semantics. Since polysemy is often grounded in encyclopedic knowledge, licensed by states of affairs in the world (e.g., we call some drinks by the name of plants or fruits from which they are made), qualia structure relations may be an indirect proof of polysemy, or can at least be used as evidence of the fact that two senses of a word are not unrelated.

Considering that the present experiment is limited to nouns, PSC uses an inventory of 157 ontological types for nominal usems; a much larger number with respect to the 39 Corelex types of the previous experiment. Corelex basic types are derived from WordNet, and most of them are mappable onto one or more SIMPLE types. So for instance Corelex type ART can map onto ARTIFACT, ARTIFACTUAL_AREA, ARTIFACTUAL_DRINK, ARTIFACTUAL_MATERIAL, ARTIFACT_FOOD in SIMPLE. Notice that SIMPLE types have different levels of specificity, which means that usems can be typed at different levels of the SIMPLE ontology. At the same time, the ontology specifies only distinctions that are linguistically relevant; in this example for instance ARTIFACTUAL_AREAS, ARTIFACTUAL_DRINKS,

[5] Again, we are aware that categorizing words into homonymous/polysemous by using this index is still an oversimplification; however, it can be useful in practice to indicate the *systematicity* of the *predominant ambiguity type* of a word.

[6] http://www.ilc.cnr.it/clips/Ontology.htm

ARTIFACTUAL_MATERIALS and ARTIFACT_FOOD receive a distinct type, while other artifacts are left underspecified. Because of its finer-grained articulation relative to CoreLex, the use of the SIMPLE Ontology might, in principle, help avoid some of the problems encountered in the previous experiment on English (Utt and Padó, 2011) which were linked to the coarse-grained nature of the ontology.
SIMPLE types were used in the construction of PSC, in particular to define templates that guided the lexicographers in the identification of the semantic properties (especially the qualia structure, but also polysemy alternations) of an entry. Thus, usems belonging to the same type are expected to show similar semantic and lexical relations.

5. Experimental setup

The general procedure – analogous to Utt and Padó (2011) – consists in an iterative evaluation of BAs found in the resource over a gold standard of lemmas, in order to determine the optimal value of N in equation (1). If we rank all BAs found in PSC from the most frequent to the least frequent, N represents the rank of each BA when ordered by descending frequency, taking into account the fact that two BAs with the same frequency have the same rank. The procedure induces the optimal value of N, such that all the most frequent BAs up to the Nth rank are to be considered regular polysemies. The induction of this optimal threshold is achieved by evaluating the frequency index of a set of known polysemous and homonymous words.
The algorithm is as follows:

1. Extract from PSC all lexically ambiguous nouns (i.e., all nouns with more than one usem).

2. For each sense, retrieve the respective ontological type (e.g., HUMAN, INSTITUTION, etc.), and add each BA pair to a list of attested BAs with their counts. Table 1 lists some of the most frequent BAs extracted from PSC and some of the lemmas in which they occur:

BA	**examples**
Information# Semiotic_artifact	targa, quotidiano, scontrino,
Language#People	italiano, sloveno, cinese,
Instrument#Profession	sagomatrice, mietitrice, tirabozze,
Body_Part#Part	braccio, palma, piede,
Instrument#Part	rotella, spina, ventola,
Amount#Container	barile, bustina, sacco,
Building#Human_Group	municipio, redazione, tribunale, ..
Earth_animal#Human	coniglio, formica, maiale,
Agent_of_persistent_activity# Profession	violinista, suonatore, cuoco,
...	...

Table 1: Some basic ambiguities found in PSC

   Overall 2198 BAs are found (a significantly larger number than the WordNet-attested Corelex BAs, 663).

3. Organize the BAs into frequency bins (FBs) (54 compared to the 39 in Utt and Padó (2011)), ranked from the most frequent to the least frequent. As shown in Table 4, the top ranked FB contains just one BA, namely INFORMATION#SEMIOTIC_ARTIFACT with frequency 218, FB 8 contains two BAs; and the last FB, not shown in the table, contains 1089 BAs, all with frequency 1.

4. Take a list of (proto-)typically homonymous and polysemous lemmas, with their usems and BAs.

5. Create an empty set and iteratively add at each run the content of the next FB, starting from the most frequent down; each time calculating the polysemy index using the content of the set as the list of regular polysemies. So, at the 1st run the set of regular polysemies will contain only one BA (INFORMATION#SEMIOTIC_ARTIFACT), namely the content of FB 1; at the 2nd run the set will contain two BAs (INFORMATION#SEMIOTIC_ARTIFACT, LANGUAGE#PEOPLE), namely FB 1+FB 2; etc.

6. During each such run, perform the Mann-Whitney U test in order to measure how well the polysemous lemmas can be distinguished from the homonyms. The U statistic is obtained by counting (for all possible pairs of homonymous/polysemous word pairs) how many are ranked correctly, that is how many times the following inequality holds:

$$\pi_N(poly) > \pi_N(hom) \quad (2)$$

7. After final run, find the optimal value for N, such that U is maximal.

We ran two experiments; in the first case, the selection of seed lemmas is done using information encoded in PSC and in the other case the lemmas are defined independently of the resource. Thus the first experiment can be seen as an internal validation of the consistency of the PSC encoded regular polysemies, while the second provides an external validation.

5.1. Experiment 1 - internal validation

In this experiment, the goal was to exploit the explicitly encoded information about polysemy in PSC to perform the induction of the optimal threshold. In order to do this, the list of (proto-)typical polysemous and homonymous lemmas (step 4 in section 5 above) is selected from PSC, in an unsupervised, purely automatic way:

- Polysemous lemmas are chosen from among those whose senses are all linked by explicitly encoded polysemy relations. Those with the higher number of senses are privileged.

- Homonymous lemmas are chosen taking those with high numbers of usems, and a near-zero number of relations between them. In this case, we exclude not only polysemy, but also qualia and metaphor relations; the idea is to obtain lemmas whose senses are not or are very distantly semantically related to one another.

As in Utt and Padó (2011), we create a list of 48 lemmas (reported in Table 2).
Using these lemmas, the number of pairwise comparisons between homonymous and polysemous lemmas in the Mann-Whitney test is 576.
As shown in Figure 1, the maximum value of the U statistic reached is 566, and is obtained when adding FB 27; at this

homonyms	corrente, area, amore, forma, metodo, ariete, colombina, varietà, centro, capo, piumino, ala, acquario, croce, calore, scambio, stampo, disegno, cucina, blocco, punto, marcia, base, unità
polysemous	chinotto, contralto, brefotrofio, cacao, ginnasio, anice, conservatorio, elementare, ospedale, dogana, orfanotrofio, convitto, ambasciata, ateneo, senato, politecnico, bergamotto, patente, soprano, baritono, mezzosoprano, università, accademia, vetreria

Table 2: List of automatically selected lemmas from PSC, prototypical for polysemy and homonymy

iteration the set of regular polysemies contains 36 BAs (see Table 4 for a list) and the frequency threshold is 28. This value is statistically highly significant – only 10 pairwise rankings are wrong.

Despite the encouraging result, we notice how the two lists of lemmas are slightly unbalanced, as shown by the progression of the U-statistic. When a frequency bin that does not contain any of the BAs instantiated by the selected set of 48 lemmas is added the U test result does not vary from the previous iteration. This leads to the presence of invariances in the U statistics, namely sequences of iterations producing the same value in the U test; such cases are recognizable as plateaus, namely horizontal lines in the plot.

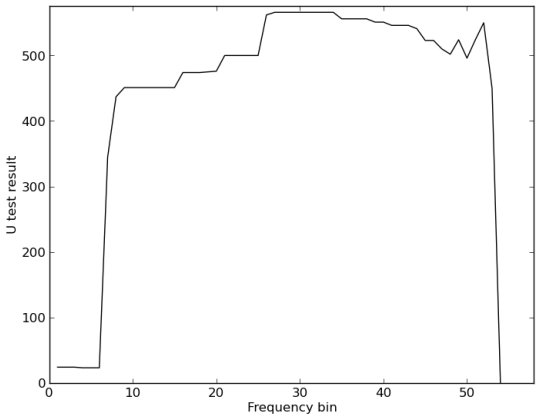

Figure 1: Results at different FBs for the U test, using a PSC based set of lemmas.

Indeed, the best result is itself located on such a plateau, between FB 27 and 34. In this experiment, the first iteration that produces the best result (the one adding FB 27 to the list of regular polisemies) is to be considered as the optimal frequency threshold for N, since the others add BAs for which no validation can be obtained from the chosen set of lemmas. At the same time, the BAs from FB 28 to 34 can be granted some "presumption of innocence", as they do not decrease the performance of the U test when added to the list.

5.2. Experiment 2 - external validation

We repeated the experiment using a manually created list of prototypical lemmas for polysemy and homonymy (Table 3) that contained a better balanced set of alternations and provided a PSC-external validation for the derived set of regular polysemies.

This list was created by drawing from the literature about polysemy and homonymy in Italian (Jezek and Quochi, 2010; Andorno, 2003; Serianni and Castelvecchi, 1988) and cross-checking both in dictionaries and PSC (to ensure the lemma is present in the resource). We ended up with a list of 30 lemmas (15 homonyms + 15 polysemous). [7].

Here the number of paiwise comparisons for the test is 255 (=15×15). In this case (cf. Figure 2), the maximum value of the U statistic obtained is 147, using FB 34 (BAs with frequency up to 21) as a cutoff and classifying 54 BAs as regular polysemies. Plotting U against N shows a curve with fewer plateaus.

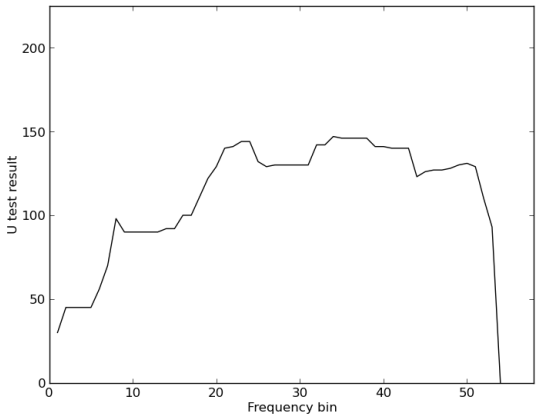

Figure 2: Results at different FBs for the U test, using a manually defined set of lemmas.

The optimal set of polysemy relations is reached by adding FB 34 (frequency threshold 21), that is at the end of the optimal-cutoff plateau of the previous induction experiment; this indicates the solidity of this threshold. Also, the maximum here is located at a single FB (no plateau) which indicates an increased discriminatory power for this setup. The ratio of 147/225 correctly classified pairs is lower than in the previous experiment, but is still significant.

Notice how, if we applied the results of the previous induction, and stopped at FB 27, we would get a not too distant result of 130/225. This means that BAs up to FB 27, that performed well with PSC chosen lemmas, do not perform as well on externally chosen ones.

This setup is similar to Utt and Padó (2011), but with fewer lemmas; results are closer to the original ones: there the authors obtained 75% correct rankings, here we obtain 65%.

[7] It is interesting to note that we had to reduce the dimension of the list with respect to the original experiment, as it seems that Italian has less cases of clear homonymy than English.

homonyms	caccia, miglio, tasso, lira, piano, calcolo, boa, banda, volta, canto, borsa, zecca, razza, botte, riso
polysemous	rosa, noce, caffè, porta, scuola, città, re, bottiglia, coniglio, libro, giornale, basilico, italiano, marmo, fotografo

Table 3: List of manually selected prototypical lemmas for polysemy and homonymy

6. Discussion

Both experiment 1 and 2 set the 'polysemy cutoff' around the same interval; this discussion takes into account the results of both, whose combined results define two lists: one up to FB 27 (36 most frequent BAs, up to frequency 28), namely up to the double line in Table 4 and one up to FB 34 (56 most frequent BAs, up to frequency 21), namely the whole Table 4.

From the point of view of the U-statistics, an interesting observation is that both graphs show a big fall in the performance only with the addition of the last two FBs (with frequency 1 and 2, that obviously contain large numbers of BAs). Instead, up to the third last FB (freq. 3) performance remains in both cases close to the maximum. In other words, it suffices that a BA is present in three different lemmas in order to make it a plausible candidate for regular polysemy, in this scenario. Indeed, in PSC only one regular polysemy relation is encoded for BAs with frequency 2 and none with frequency 1, but some polysemy relations are encoded for BAs with frequency 3 or less.

6.1. Evaluation of induced BAs

The results obtained appear partly consistent with the PSC-encoded relations.

In Table 4 the fourth column lists all possible relations between usems that are encoded in PSC for that BA. The relations are listed in the format *relation-name#relation-category*, where the relation category can be one of the 4 qualia (Constitutive, Formal, Telic, Agentive), Metaphor or Polysemy, while *relation-name* better specifies the type of relation (the Metaphor category has no subspecifications).

When comparing the induced results with PSC, four cases can be recognized:

A) a BA is matched by just one encoded polysemy relation

B) more than one encoded polysemy corresponds to one BA

C) no polysemy relation is present but at least another lexical relation (metaphor or derivation) is present

D) only qualia relations exist between the alternating usems of a lemma that expresses a BA

E) no relation at all is encoded in PSC for a BA.

In all cases but (A) it is obviously possible that a regular polysemy is involved that had been not foreseen in the design of PSC. In Table 5 a list of all cases is given.

(A) represents the perfect validation, and occurs in 15 cases over 36 for the FB 27 set; as for the FB 34 set, the 18 added BAs contain only two more such cases. This shows how PSC lexicographers have worked on the whole quite consistently, first by encoding frequent BAs into explicit polysemy relations, and then by using such relations systematically in all relevant usem pairs. Classic polysemy cases are to be found here, such as *PolysemySemioticartifact-Information* (e.g., 'letter', 'newspaper'); *PolysemyPlant-Flower*; etc. The presence of qualia relations, often Constitutive, does not impact on the goodness of this result, but shows how some polysemies may be due to meronymic sense shifts.

(B) is a rare phenomenon, with only three occurrences all located above FB 27. Such a case is the presence of *PolysemyHumanGroup-Institution* alongside with *PolysemyBuilding-Institution*, for BA BUILDING#INSTITUTION; on closer look such cases seem the product of slight inconsistencies from the part of the lexicographer, and the incorrect polysemy relation occurs in just one lemma out of all those showing that BA.

(C) cases are more interesting, since they illustrate phenomena that may cast a doubt on the frequency based definition of polysemy followed in the present work. Here some very frequent BAs are classified by the lexicographer in terms of zero derivation (such as instrument *violino*, 'violin' INSTRUMENT, used for the PROFESSION, violinist) or of metaphorical extension (such as *coniglio*, 'rabbit', for a cowardly person). Such cases are frequent, probably even semi-productive, but lack the regularity that characterizes systematic polysemy.

(D) cases occur rarely in BAs above both cut-offs (4,5 times), and the qualia relations listed occur very rarely among the corresponding lemmas. Such lemmas, though not strictly polysemous, represent instances of semanticized metaphoric extension of the sort that may qualify for formal encoding with the *metaphor* relation; so for instance *spada*, 'sword', has a sense typed under AGENT_OF_TEMPORARY_ACTIVITY to indicate uses such as *He is a good sword* meaning 'He is a good swordsman'.

(E) cases require careful analysis, since they are the most problematic outcome. First of all notice how their number considerably increases when adding FBs 28 to 34; again this is no surprise given that the FB 27 threshold is the outcome of the internal validation experiment. Concerning the 7 cases occurring up to FB 27, some of them seem to be the outcome of semi-productive phenomena, despite the lack of lexicographic encoding. So for instance, BODY_PART#PART, with frequency 101, captures the fact that parts of artifacts (e.g. machines, ships, ...) are often denoted in Italian by using words for body parts (such as in *braccio*, used for: 'person's arm', 'gramophone's arm', 'edifice's wing'); PSC lexicographers did not define an explicit relation for such alternations, but they seems more cases of metaphorical extension than of regular polysemy.

Other (E) alternations instead show clearly related senses and a higher level of systematicity. Such is the case with

FB	BA	freq.	PSC semantic relations encoded among usems with this BA
1	Information#Semiotic_artifact	218	Contains#Constitutive, Isin#Constitutive, PolysemySemioticartifact-Information#Polysemy
2	Language#People	174	PolysemyPeople-Language#Polysemy
3	Instrument#Profession	108	NounNoun#Derivational, Usedby#Telic , Uses#Constitutive
4	Body_Part#Part	101	
5	Instrument#Part	98	Isapartof#Constitutive
6	Amount#Container	79	PolysemyContainer-Amount#Polysemy
7	Building#Human_Group	74	PolysemyHumanGroup-Building#Polysemy, PolysemyBuilding-Institution#Polysemy, PolysemyHumanGroup-Institution#Polysemy
8	Earth_animal#Human	73	Metaphor
8	Agent_of_persistent_activity# Profession	73	PolysemyAgentofpersistentactivity-Profession#Polysemy
9	Building#Institution	63	PolysemyBuilding-Institution#Polysemy, PolysemyHumanGroup-Institution#Polysemy
10	Agent_of_temporary_activity# Profession	59	
11	Substance_food#Water_animal	58	PolysemyAnimal-Food#Polysemy
12	Agent_of_temporary_activity# Instrument	57	Uses#Constitutive
13	Act#Psych_property	54	
14	Human_Group#Institution	53	PolysemyHumanGroup-Institution#Polysemy
15	Human#Profession	52	Metaphor
16	Plant#Vegetable	49	PolysemyPlant-Vegetable#Polysemy, Producedby#Constitutive, Produces#Constitutive
17	Human#Representation	48	Metaphor
18	Human#Instrument	45	Metaphor
18	Flower#Plant	45	PolysemyPlant-Flower#Polysemy, Producedby#Constitutive, Produces#Constitutive
19	Convention#Semiotic_artifact	44	Isin#Constitutive, PolysemyConvention-Semioticartifact#Polysemy
20	Natural_substance#Plant	39	PolysemyPlant-Substance#Polysemy, Producedby#Constitutive, Madeof#Constitutive, Produces#Constitutive
21	Body_Part#Instrument	38	
21	Building#Group	38	Concerns#Constitutive, Hasaspart#Constitutive
21	Artifactual_material#Earth_animal	38	Derivedfrom#Agentive, PolysemyAnimal-Material#Polysemy
22	Human_Group#Part	35	
22	Geopolitical_location#Human_Group	35	PolysemyHumanGroup-GeopoliticalLocation#Polysemy
23	Earth_animal#Substance_food	34	PolysemyAnimal-Food#Polysemy
24	Human#Social_status	33	Metaphor
25	Domain#Symbolic_Creation	30	
25	Color#Natural_substance	30	Hasascolour#Constitutive, PolysemySubstance-Colour#Polysemy
26	Fruit#Plant	29	PolysemyPlant-Vegetable#Polysemy, PolysemyPlant-Fruit#Polysemy, Producedby#Constitutive, Produces#Constitutive
26	Human#Ideo	29	DeadjectivalNoun#Derivational, Metaphor
26	Building#Domain	29	Concerns#Constitutive
26	Clothing#Instrument	29	
27	Artifactual_drink#Plant	28	PolysemyPlant-ArtifactualDrink#Polysemy, Madeof#Constitutive
28	Act#Quality	27	
28	Physical_property#Quality	27	
28	Constitutive#Instrument	27	Metaphor
29	Psych_property#Quality	26	
30	Air_animal#Human	25	Metaphor
30	Container#Part	25	Isapartof#Constitutive, Hasaspart#Constitutive
30	Constitutive#Part	25	
31	Building#Part	24	
32	Human#Substance_food	23	
32	Container#Instrument	23	
32	Artwork#Domain	23	
32	Agent_of_temporary_activity#Human	23	Metaphor
32	Instrument#Vehicle	23	
32	Flavouring#Plant	23	Producedby#Constitutive, PolysemyPlant-Flavouring#Polysemy, Produces#Constitutive
33	Cause_Change_of_State#Change_of_State	22	
33	Part#Purpose_Act	22	
34	Color#Plant	21	PolysemyVegetalEntity-Colour#Polysemy
34	Group#Human_Group	21	Metaphor

Table 4: BAs selected by first (double line at FB 27) and second experiment (FB 34) and PSC relations encoded for them.

AGENT_OF_PERSISTENT_ACTIVITY#PROFESSION, typical of lemmas such as *pianista*, 'pianist', denoting both someone who plays piano professionally and someone who plays piano regularly, but as an amateur. Another such case is ACT#PSYCH_PROPERTY, with lemmas such as *idiozia*, 'silliness', once listed as the property of associated with being an idiot and then with the act of being idiotic. Such alternations are rarely listed among the known polysemy alternations, and are the product of the semantic richness of PSC and of the SIMPLE ontology, that distinguishes shades of meaning that are normally not taken into account in other resources. At the same time, within the context of PSC, they are quite systematic and may be considered for an explicit encoding.

Finally, some (E) cases are somewhat epiphenomenal: so for instance HUMAN#SUBSTANCE_FOOD is not the product of some macabre cannibalistic practice, but the result of the fact that some animals, typically those familiar animals that are used for food, are also used to metaphorically define properties of humans, such as *pig*, *chicken* and *goat*. In this case, there is a pivotal usem (the ANIMAL one) that is linked to the other two by separate alternations (ANIMAL#HUMAN and ANIMAL#SUBSTANCE_FOOD), producing an indirect alternation (HUMAN#SUBSTANCE_FOOD).

	A	B	C	D	E	TOT
FB 27	15	3	7	4	7	36
FB 34	17	3	11	5	18	54

Table 5: Comparison between induced BAs and lexical semantic relations in PSC, for both induced thresholds.

Interestingly, we noticed that most of the critical, or less clear, cases presented above tend to be found between FB27 and FB34[8].

6.2. Finding gaps in PSC

In order to extract possible gaps in PSC, the polysemy index is calculated for all lemmas. Then the presence of relations among usems is assessed. Finally the two results are compared, to find cases where the polysemy index is high but the resource lists few or no relations. The most restrictive case is chosen, with FB 27 as cut-off (so fewer BAs are counted as polysemy relations), and a matching algorithm extracting only names with $\pi_{27} = 1$ and no relation whatsoever among its usems (including non polysemy relations, such as qualia).

The test was carried out on the 4905 polysemous nouns in PSC and produced 632 cases (of lemmas with high polysemy index, but no corresponding relation).

One of the most striking examples of a gap involves the LANGUAGE#PEOPLE BA, which is not only identified as polysemic by our bottom-up methodology, but also explicitly encoded in PSC by means of the *PolysemyPeopleLanguage* relation, typically used in cases such as *italiano* ('Italian') and *francese* ('French'). Nevertheless, out of the 174 occurrences of this BA, 55 are not explicitly encoded. For instance the word *miceneo* ('Mycenaean'), which shows a polysemy index of 1.0 at FB 27, has no encoded polysemy relation between its two usems.

The same is true for:
INFORMATION#SEMIOTIC_ARTIFACT (13 cases),
BUILDING#INSTITUTION (4 cases),
COLOR#NATURAL_SUBSTANCE (2 cases),
NATURAL_SUBSTANCE#PLANT (6 cases),
ARTIFACTUAL_MATERIAL#EARTH_ANIMAL (38 cases),
HUMAN_GROUP#INSTITUTION (2 cases).

Other frequent cases - such as *arpista* ('harpist'), *disegnatore* ('graphic designer'), *speleologo* ('speleologist'), *motociclista* ('motorcyclist') are due to the fact that some alternations are absent in PSC despite the fact that they occur quite systematically (the aforementioned AGENT_OF_TEMPORARY_ACTIVITY#PROFESSION and AGENT_OF_PERSISTENT_ACTIVITY#PROFESSION).

In such cases, a new polysemy relation could first be introduced and then applied to all instances in the resource.

Finally, some cases are hard to amend, or even to define as gaps in the lexicon. They are instances of alternations that have been discussed previously as being frequent but not prototypically polysemous. Thus, for instance, a *rispenditore*, 'answerer', can be a person that has the task of answering or an answering machine thus alternating between AGENT_OF_TEMPORARY_ACTIVITY and INSTRUMENT alternations; the relationship between the two senses is hard to pin down, even in terms of derivation, metaphoric extension or qualia structure (the person does not use the machine, nor build the machine – he or she just performs the same job); thus no encoded relation is present. Nevertheless, the index clearly records the fact that the two senses are related and that this is no case of homonymy. In such cases, it may be worthwhile to explicitly encode this underspecified relationship between the senses.

7. Conclusions

The obtained results can be viewed as a means for checking the consistency of a lexicon; a list of typical polysemous/homonymous lemmas is used to induce a polysemy threshold for basic type alternations, and later the threshold is projected onto other lemmas by calculating a polysemy index. This can lead to the discovery of new basic ambiguities that may need to be encoded with an explicit polysemy relation.

At the same time, the comparison of the induced set of polysemic alternations to encoded lexicographic knowledge leads to more general considerations concerning the proposed methodology. The FB 27 threshold does seem to identify a strong group of known regular polysemies, but at the same time it also promotes to regular polysemy slightly less prototypical cases.

This seems to imply that frequency alone is not a sufficient enough a criterion to define *systematic* polysemy. The proposed methodology seems to be more reliable in distinguishing any kind of polysemy alternation between related senses. Indeed, hardly any BAs above the threshold show up in nouns with totally unrelated senses. When analyzing the inconsistencies between the PSC encoded relations among senses and the polysemy index at FB 27, the results

[8] A more thorough analysis of such cases however is required, which we reserve to future work.

show that most words with a high polysemy index have related senses, even when the relationship is hard to define, thus excluding homonymy.

Future work will further investigate ways to distinguish systematic from unsystematic polysemy as well as from homonymy by exploiting the rich set of ontological and semantic relations of PSC. More specifically, the qualia structure links between usems may be used to identify those qualia relations between senses that constitute the ontological trigger for prototypical polysemic alternations.

8. References

Andorno, C. (2003). *La grammatica italiana, Volume 15; Volume 50*. B. Mondadori.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 142:5–32.

Copestake, A. and Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

Jezek, E. and Quochi, V. (2010). Capturing Coercions in Texts: a First Annotation Exercise. In *LREC*.

Landau, S. I. (1984). *Dictionaries: The Art and Craft of Lexicography*. Charles Scribner's Sons, New York.

Leech, G. N. (1974). *Semantics*. Penguin, Harmondsworth.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowski, A., Peters, I., Peters, W., Ruimy, N., Villegas, M., and Zampolli, A. (2000). Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Lyons, J. (1977). *Semantics. Vol 2*. Cambridge University Press, Cambridge.

Malmberg, S. (1988). On regular polysemy in swedish. *Studies in Computer-aided Lexicography*.

Ndlovu, E. and Sayi, S. (2010). The treatment of polysemy and homonymy in monolingual general-purpose dictionaries with special reference to "isichazamazwi sesindebele". *Lexikos*, 20:351–370.

Nunberg, G. and Zaenen, A. (1992). Systematic polysemy in lexicology and lexicography. In *Proceedings of Euralex II*, pages 387–395, Tampere, Finland.

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics*, 12(2):109–132.

Palmer, F. R. (1981). *Semantics*. Cambridge University Press, Cambridge.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge MA.

Ruimy, N., Corazzari, O., Gola, E., Spanu, A., Calzolari, N., and Zampolli, A. (1998). The european le-parole project: The italian syntactic lexicon. In *Proceedings of the First International Conference on Language resources and Evaluation*, pages 241–248.

Saeed, J. I. (1997). *Semantics*. Blackwell Publishers, Oxford.

Serianni, L. and Castelvecchi, A. (1988). *Grammatica italiana:*. Nostra lingua. UTET.

Toral, A. and Monachini, M. (2007). Simple-owl: a generative lexicon ontology for nlp and the semantic web. In *Workshop of Cooperative Construction of Linguistic Knowledge Bases, 10th Congress of Italian Association for Artificial Intelligence.*

Utt, J. and Padó, S. (2011). Ontology-based distinction between polysemy and homonymy. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, Oxford, UK.

Zgusta, L. (1971). *Manual of Lexicography*. Mouton, The Hague/Paris.