10.12871/CLICIT2014134

Polysemy alternations extraction using the PAROLE SIMPLE CLIPS Italian lexicon.

Francesca Frontini **Valeria Quochi** **Monica Monachini**
Istituto di Linguistica Computazionale "A.Zampolli" CNR Pisa
name.surname@ilc.cnr.it

Abstract

English. This paper presents the results of an experiment of polysemy alternations induction from a lexicon (Utt and Padó, 2011; Frontini et al., 2014), discussing the results and proposing an amendment in the original algorithm.

Italiano. *Questo articolo presenta i risultati di un esperimento di induzione di alternamze polisemiche regolari (Utt and Padó, 2011; Frontini et al., 2014), discutendone i risultati e proponendo una modifica all'originale procedura.*

1 Introduction

The various different senses of polysemic words do not always stand to each other in the same way. Some senses group together along certain dimensions of meaning while others stand clearly apart. Machine readable dictionaries have in the past used coarse grained sense distinctions but often without any explicit indication as to whether these senses were related or not. Most significantly, few machine readable dictionaries explicitly encode systematic alternations.

In Utt and Padó (2011) a methodology is described for deriving systematic alternations of senses from WordNet. In Frontini et al. (2014) the work was carried out for Italian using the PAROLE SIMPLE CLIPS lexicon (PSC) (Lenci et al., 2000), a lexical resource that contains a rich set of explicit lexical and semantic relations. The purpose of the latter work was to test the methodology of the former work against the inventory of regular polysemy relations already encoded in the PSC semantic layer. It is important to notice that this was not possible in the original experiment, as WordNet does not contain such information.

The result of the work done on PSC shows how the original methodology can be useful in testing the consistency of encoded polysemies and in finding gaps in individual lexical entries. At the same time the methodology is not infallible especially in distinguishing type alternations that frequently occur in the lexicon due to systematic polysemy from other alternations that are produced by metaphoric extensions, derivation or other non systematic sense shifting phenomena.

In this paper we shall briefly outline the problem of lexical ambiguity; then describe the procedure of type induction carried out in the previous experiments, discussing the most problematic results; finally we will propose a change in the original methodology that seems more promising in capturing the essence of systematic polysemy.

2 Theoretical background on lexical ambiguity

Two main types of lexical ambiguity are usually distinguished, homonymy and polysemy.

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.

The notion of polysemy in contrast foresees a commonality of meaning that is shared between the different senses of the same word. Polysemy has received ample treatment in the literature (Apresjan, 1974; Nunberg and Zaenen, 1992; Copestake and Briscoe, 1995; Nunberg, 1995; Palmer, 1981). Three main types can be identified. **Regular (or logical polysemy)**: Words with two, or 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 or ontological categories to which the senses of a lemma belong (Palmer, 1981; Pustejovsky, 1995). Well known cases of regular alternations are ANIMAL−FOOD, BUILDING−INSTITUTION. These systematic meaning alternations are generally salient on conceptual grounds, common to (several) other words[1], and usually derivable by metonymic sense shifts.

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)

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.

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.

3 Previous experiments

We refer to Utt and Padó (2011) and Frontini et al. (2014) for a precise description of the experiment on English and Italian respectively and of the induction algorithm. Here an intuitive outline is given. If we consider a lemma and all of its senses, each possible sense can be labeled 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 labeled 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 is the product of ANIMAL >FOOD regular polysemy.

The induction algorithm proposed essentially derives the complete list of BAs from a given lexicon by extracting all type alternations occurring within polysemous lemmas (nouns in our case), and ranks them per descending frequency. The assumption is that the most frequent BAs will be polysemous, whereas the less frequent ones will be occasional. The optimal frequency threshold N for a BA to be classified as a regular polysemy is induced by testing it against a set of known homonymous and polysemous seed lemmas. The correct threshold is the one that correctly separates typically homonyms from polysemous words.

In Frontini et al. (2014) we run two experiments with two different sets of seeds and derived two frequency thresholds (≥ 28 and ≥ 21 respectively), identifying a set of overall 36 and 54 Basic Alternations that can be considered polysemous (see the cited paper for the difference between the two thresholds). In the present paper we shall refer mostly to the frequency threshold ≥ 21, which was derived by strictly following the methodology proposed by Utt and Padó (2011), namely using a set of prototypically polysemous/homonymous lemmas drawn form the literature.

In Frontini et al. (2014) we report on the results above the first and second threshold. Each induced BA is compared with all possible relations that are encoded in PSC between senses of words exhibiting that BA. Relations encoded among senses in PSC are of two types: Lexical relations (such as Polysemy itself, Metaphor, Derivation) or Qualia relations (Constitutive, Formal, Telic, Agentive), following the generative lexicon theory (Puste-

[1] Of course some exceptions are possible, e.g. *cow/beef*. (Nunberg, 1995; Copestake and Briscoe, 1995)

jovsky, 1995).

When comparing the induced results with PSC, four cases can be recognized [2]:

A) a BA is matched by one or more polysemy relations

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 not been foreseen in the design of PSC. In the first line of Table 1 the results for the original experiment are given.

(A) represents the perfect validation. 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.

(C) cases are the more interesting ones, 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, 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'.

[2]Case A in the present paper merges cases A and B of the previous one; the original labelling for the other cases is maintained for comparison.

(E) cases require careful analysis, since they are the most problematic outcome. Some of them seem to be the result 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, as they seem 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 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 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).

The conclusion drawn from this experiment was 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.

4 New experiment and preliminary conclusion

While distinguishing when two senses are totally unrelated may indeed be very useful, the original goal of this research was to be able to automatically detect regular polysemy alternations. In this new experiment we then try to see if the original methodology can be improved in order to make it more capable to single out systematic polysemy, which is characterised by productivity and ontological grounding.

The ontological grounding of polysemy can be assessed in resources such as PSC by checking the qualia relations; indeed many of the officially encoded polysemies in PSC co-occur with qualia relations. Nevertheless this methodology can hardly be automatized or applied to other resources such as WordNet that lack qualia information. As for the productivity, it is clearly related to the directionality of the polysemy rule. If the directionality is from type A to type B we can presume that all words that have a sense of type A can be also used in a sense of type B, but not vice versa. So if the rule is "Animal to Food", then all words for Animal should also have the Food sense, but not vice versa. So *crockodile* can denote food in some contexts, but *spaghetti* cannot be used to refer to an animal.

In a methodology such as the one proposed it is hard to retrieve directionality from polysemy rules, since lexicons are rarely exhaustive. Nevertheless it may be possible to indirectly asses the systematicity of the type alternation by comparing the frequency of the BA with the one of each type separately. An efficient way to treat this problem is to consider measuring the association strength of the two types by using Pointwise Mutual Information, following what has been previously proposed in Tomuro (1998). PMI assigns the maximum value to pairs that occur only together, and in general gets higher values if at least one of the two elements occurs with the other more frequently than alone. It is calculated as:

$$PMI(t1, t2) = \log \frac{\frac{f(t1,t2)}{N}}{\frac{f(t1)}{N} X \frac{f(t2)}{N}} \qquad (1)$$

where t1 and t2 are the number of lemmas in which of each of the two ontological types of a BA occur overall, and (t1,t2) is the number of lemmas in which they occur together. Taking into account the tendency of PMI to promote hapaxes, a raw frequency filter ≥ 5 for co-occurrencies values was implemented. We thus rank the BAs in PSC using descending PMI instead of raw frequency, then we induce the optimal PMI threshold following the standard procedure, using the same set of 12 + 12 typically polysemous/homonymoys lemmas drawn from the literature, and comparing the results. The second line of Table 1 shows the cases obtained from this new experiment, while table 4 presents the complete list. The number of BA induced with the two ranking systems is comparable (49 for PMI vs 54 for raw frequency).

	A	C	D	E	TOT
F $>$ 21	20	11	5	18	54
PMI $>$ 1.8	24	1	8	16	49

Table 1: Comparison between induced BAs and lexical semantic relations in PSC, for both induced thresholds.

First results seem promising. Most significantly PMI ranking promotes only one C case above the threshold, vs 11. LOCATION#OPENING, is indeed a Metaphor occurring only 8 times in PSC, in cases such as "topaia" (rathole) that can be used to metaphorically refer to human abodes in very unflattering terms.

This seems to signify that PMI ranking is more effective in demoting cases unsystematic polysemy. Remarkably PMI ranking demotes one of the most problematic and frequent of the previously discussed BA, BODY_PART#PART, under the threshold while promoting a larger number of the encoded polysemies to the top. In the first 18 positions we find only one gap at position 8 and it turns out that this BA - CONVENTION#MONEY - is actually a good candidate for systematic polysemy, as MONEY is both an artifact and a human convention.

To conclude, such preliminary results actually seem to confirm the hypothesis that measuring the association strength between types, rather than the frequency of their cooccurrence, is useful to capture the systematicity of an alternation.

In future work it may be interesting to test ranking by other association measures (such as Log Likelihood) and with different filternigs. Finally, the original experiment may be repeated on both Italian and English WordNets in order to evaluate the new method on the original lexical resource.

References

Jurij D. Apresjan. 1974. Regular Polysemy. *Linguistics*, 142:5–32.

Ann Copestake and Ted Briscoe. 1995. Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, 12:15–67.

Francesca Frontini, Valeria Quochi, Sebastian Padó, Monica Monachini, and Jason Utt. 2014. Polysemy Index for Nouns: an Experiment on Italian using the PAROLE SIMPLE CLIPS Lexical Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2955–2963, Reykjavik, Iceland.

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

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

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4):249–263.

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

E. Ndlovu and S. Sayi. 2010. The Treatment of Polysemy and Homonymy in Monolingual General-purpose Dictionaries with Special Reference to "Isichazamazwi SesiNdebele". *Lexikos*, 20:351–370.

Geoff Nunberg and Annie Zaenen. 1992. Systematic polysemy in lexicology and lexicography. In *Proceedings of Euralex II*, pages 387–395, Tampere, Finland.

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

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

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge MA.

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

Noriko Tomuro. 1998. Semi-automatic induction of systematic polysemy from WordNet. In *Proceedings ACL-98 Workshop on the Use of WordNet in NLP*.

Jason Utt and Sebastian Padó. 2011. Ontology-based Distinction between Polysemy and Homonymy. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, Oxford, UK.

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

BA	Val.	freq.	PMI
Substance_food#Water_animal	A	58	4.66
Flower#Plant	A	45	4.40
Information#Semiotic_artifact	A	218	4.26
Plant#Vegetable	A	49	4.16
Flavouring#Plant	A	23	4.13
Color#Flower	A	7	3.94
Color#Fruit	A	9	3.85
Convention#Money	D	16	3.82
Fruit#Plant	A	29	3.62
Building#Institution	A	63	3.39
Amount#Container	A	79	3.39
Language#People	A	174	3.35
Earth_animal#Substance_food	A	34	3.33
Color#Vegetable	A	5	3.27
Convention#Semiotic_artifact	A	44	3.09
Artifactual_drink#Plant	A	28	3.04
Human_Group#Institution	A	53	3.01
Color#Natural_substance	A	30	2.98
Area#VegetalEntity	D	6	2.92
Concrete_Entity#Transaction	D	5	2.83
Cause_Change_of_State#Material	E	14	2.82
Artifactual_material#Earth_animal	A	38	2.80
Color#Plant	A	21	2.75
Air_animal#Substance_food	A	12	2.71
Artwork#Color	E	6	2.71
Location#Opening	C	8	2.71
Copy_Creation#Semiotic_artifact	D	5	2.62
Cause_Constitutive_Change# Constitutive_Change	E	5	2.60
Area#Artifactual_area	E	6	2.42
Artwork#Symbolic_Creation	D	9	2.40
Act#Psych_property	E	54	2.40
Artifactual_material#Substance_food	E	10	2.39
Food#Time	D	5	2.37
Amount#D_3_Location	E	8	2.30
Constitutive#Shape	D	7	2.22
Artifactual_material#Artwork	A	8	2.17
Convention#Institution	E	7	2.16
Plant#VegetalEntity	D	10	2.13
Convention#Time	E	15	2.11
Amount#Transaction	E	7	2.05
Time#Unit_of_measurement	E	7	2.04
Cause_Change#Change	E	5	1.98
Artifact#Artifact_Food	E	6	1.96
Abstract_Entity#Metalanguage	E	8	1.95
Artifactual_material#Color	E	5	1.89
Building#Human_Group	A	74	1.86
Natural_substance#Plant	A	39	1.85
Number#Time	E	6	1.84
Convention#Information	A	13	1.83

Table 2: BA induced using PMI as ranking method; letters represent the validation against PSC encoded relations. The order between the two types for each BA is purely alphabetical.