Annotation of regular polysemy and underspecification

Héctor Martínez Alonso,
Bolette Sandford Pedersen
University of Copenhagen
Copenhagen (Denmark)
alonso@hum.ku.dk, bsp@hum.ku.dk

Núria Bel
Universitat Pompeu Fabra
Barcelona (Spain)
nuria.bel@upf.edu

Abstract

We present the result of an annotation task on regular polysemy for a series of semantic classes or *dot types* in English, Danish and Spanish. This article describes the annotation process, the results in terms of inter-encoder agreement, and the sense distributions obtained with two methods: majority voting with a theory-compliant backoff strategy, and MACE, an unsupervised system to choose the most likely sense from all the annotations.

1 Introduction

This article shows the annotation task of a corpus in English, Danish and Spanish for regular polysemy. Regular polysemy (Apresjan, 1974; Pustejovsky, 1995; Briscoe et al., 1995; Nunberg, 1995) has received a lot of attention in computational linguistics (Boleda et al., 2012; Rumshisky et al., 2007; Shutova, 2009). The lack of available sense-annotated gold standards with underspecification is a limitation for NLP applications that rely on dot types[1] (Rumshisky et al., 2007; Poibeau, 2006; Pustejovsky et al., 2009).

Our goal is to obtain human-annotated corpus data to study regular polysemy and to detect it in an automatic manner. We have collected a corpus of annotated examples in English, Danish and Spanish to study the alternation between senses and the cases of underspecification, including a contrastive study between languages. Here we describe the annotation process, its results in terms of inter-encoder agreement, and the sense distributions obtained with two methods: majority voting with a theory-compliant backoff strategy and, MACE an unsupervised system to choose the most likely sense from all the annotations.

[1]The corpus is freely available at http://metashare.cst.dk/repository/search/?q=regular+polysemy

2 Regular polysemy

Very often a word that belongs to a semantic type, like Location, can behave as a member of another semantic type, like Organization, as shown by the following examples from the American National Corpus (Ide and Macleod, 2001) (ANC):

a) *Manuel died in exile in 1932 in England.*
b) *England was being kept busy with other concerns*
c) *England was, after all, an important wine market*

In case a), *England* refers to the English territory (Location), whereas in b) it refers arguably to England as a political entity (Organization). The third case refers to both. The ability of certain words to switch between semantic types in a predictable manner is referred to as *regular polysemy*. Unlike other forms of meaning variation caused by metaphor or homonymy, regular polysemy is considered to be caused by metonymy (Apresjan, 1974; Lapata and Lascarides, 2003). Regular polysemy is different from other forms of polysemy in that both senses can be active at the same in a predicate, which we refer to as *underspecification*. Underspecified instances can be broken down in:

1. Contextually complex: *England was, after all, an important wine market*
2. Zeugmatic, in which two mutually exclusive readings are coordinated: *England is conservative and rainy*
3. Vague, in which no contextual element enforces a reading: *The case of England is similar*

3 Choice of semantic classes

The Generative Lexicon (GL) (Pustejovsky, 1995) groups nouns with their most frequent metonymic sense in a semantic class called a *dot type*. For English, we annotate 5 dot types from the GL:

1. **Animal/Meat**: *"The chicken ran away"* vs.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 725–730,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

"the chicken was delicious".

2. **Artifact/Information** : *"The book fell"* vs. *"the book was boring"*.
3. **Container/Content**: *"The box was red"* vs. *"I hate the whole box"*.
4. **Location/Organization**: *"England is far"* vs. *"England starts a tax reform"*.
5. **Process/Result**: *"The building took months to finish"* vs. *"the building is sturdy"*.

For Danish and Spanish, we have chosen Container/Content and Location/Organization. We chose the first one because we consider it the most prototypical case of metonymy from the ones listed in the GL. We chose the second one because the metonymies in locations are a common concern for Named-Entity Recognition (Johannessen et al., 2005) and a previous area of research in metonymy resolution (Markert and Nissim, 2009).

4 Annotation Scheme

For each of the nine (five for English, two for Danish, two for Spanish) dot types, we have randomly selected 500 corpus examples. Each example consists of a sentence with a selected *headword* belonging to the corresponding dot type. In spite of a part of the annotation being made with a contrastive study in mind, no parallel text was used to avoid using translated text. For English and Danish we used freely available reference corpora (Ide and Macleod, 2001; Andersen et al., 2002) and, for Spanish, a corpus built from newswire and technical text (Vivaldi, 2009).

For most of the English examples we used the words in Rumshisky (2007), except for Location/Organization. For Danish and Spanish we translated the words from English. We expanded the lists using each language's wordnet (Pedersen et al., 2009; Gonzalez-Agirre et al., 2012) as thesaurus to make the total of occurrences reach 500 after we had removed homonyms and other forms of semantic variation outside of the purview of regular polysemy.

For Location/Organization we have used high-frequency names of geopolitical locations from each of the corpora. Many of them are corpus-specific (e.g. *Madrid* is more frequent in the Spanish corpus) but a set of words is shared: *Afghanistan, Africa, America, China, England, Europe,Germany, London*.

Every dot type has its particularities that we had to deal with. For instance, English has lexical alternatives for the meat of several common animals, like *venison* or *pork* instead of *deer* and *pig*. This lexical phenomenon does not impede metonymy for the animal names, it just makes it less likely. In order to assess this, we have included 20 examples of *cow*. The rest of the dataset consists of animal names that do not participate in this lexical alternation, like *eel*, *duck*, *chicken*, or *sardine*.

We call the first sense in the pair of metonyms that make up the dot type the *literal* sense, and the second sense the *metonymic* sense, e.g. Location is the literal sense in Location/Organization.

Each block of 500 sentences belonging to a dot type was an independent annotation subtask with an isolated description. The annotator was shown an example and had to determine whether the headword in the example had the literal, metonymic or the underspecified sense. Figure 1 shows an instance of the annotation process.

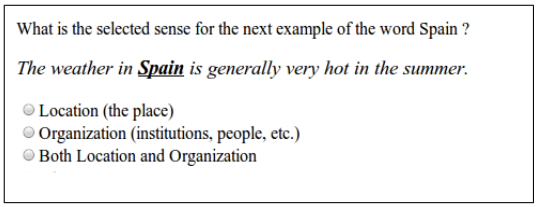

Figure 1: Screen capture for a Mechanical Turk annotation instance or HIT

This annotation scheme is designed with the intention of capturing literal, metonymic and underspecified senses, and we use an inventory of three possible answers, instead of using Markert and Nissim's (Markert and Nissim, 2002; Nissim and Markert, 2005) approach with fine-grained sense distinctions, which are potentially more difficult to annotate and resolve automatically. Markert and Nissim acknowledge a *mixed* sense they define as being literal and metonymic at the same time.

For English we used Amazon Mechanical Turk (AMT) with five annotations per example by turkers certified as Classification Masters. Using AMT provides annotations very quickly, possibly at the expense of reliability, but it has been proven suitable for sense-disambiguation task (Snow et al., 2008). Moreover, it is not possible to obtain annotations for every language using AMT. Thus, for Danish and Spanish, we obtained annotations from volunteers, most of them native or very proficient non-natives. See Table 1 for a summary of the annotation setup for each language.

After the annotation task we obtained the agree-

Language	annotators	type
Danish	3-4	volunteer
English	5	AMT
Spanish	6-7	volunteer

Table 1: Amount and type of annotators per instance for each language.

ment values shown in Table 2. The table also provides the abbreviated names of the datasets.

Dot type	$\overline{A_o} \pm \sigma$	α
eng:animeat	0.86 ± 0.24	0.69
eng:artinfo	0.48 ± 0.23	0.12
eng:contcont	0.65 ± 0.28	0.31
eng:locorg	0.72 ± 0.29	0.46
eng:procres	0.5 ± 0.24	0.10
da:contcont	0.32 ± 0.37	0.39
da:locorg	0.73 ± 0.37	0.47
spa:contcont	0.36 ± 0.3	0.42
spa:locorg	0.52 ± 0.28	0.53

Table 2: Averaged observed agreement and its standard deviation and alpha

Average observed agreement ($\overline{A_o}$) is the mean across examples for the proportion of matching senses assigned by the annotators. Krippendorff's alpha is an aggregate measure that takes chance disagreement in consideration and accounts for the replicability of an annotation scheme. There are large differences in α across datasets.

The scheme can only provide *reliable* (Artstein and Poesio, 2008) annotations ($\alpha > 0.6$) for one dot type[2]. This indicates that not all dot types are equally easy to annotate, regardless of the kind of annotator. In spite of the number and type of annotators, the Location/Organization dot type gives fairly high agreement values for a semantic task, and this behavior is consistent across languages.

5 Assigning sense by majority voting

Each example has more than one annotation and we need to determine a single sense tag for each example. However, if we assign senses by majority voting, we need a backoff strategy in case of ties.

The common practice of backing off to the most frequent sense is not valid in this scenario, where there can be a tie between the metonymic and the underspecified sense. We use a backoff that incorporates our assumption about the relations between senses, namely that the underspecified sense sits between the literal and the metonymic senses:

1. If there is a tie between the underspecified and literal senses, the sense is **literal**.
2. If there is a tie between the underspecified and metonymic sense, the sense is **metonymic**.
3. If there is a tie between the literal and metonymic sense or between all three senses, the sense is **underspecified**.

Dot type	L	M	U	V	B
eng:animeat	358	135	7	3	4
eng:artinfo	141	305	54	8	48
eng:contcont	354	120	25	0	25
eng:locorg	307	171	22	3	19
eng:procres	153	298	48	3	45
da:contcont	328	82	91	53	38
da:locorg	322	95	83	44	39
spa:contcont	291	140	69	54	15
soa:locorg	314	139	47	40	7

Table 3: **L**iteral, **M**etonymic and **U**nderspecified sense distributions, and underspecified senses broken down in **V**oting and **B**ackoff

The columns labelled L, M and U in Table 3 provide the sense distributions for each dot type. The preference for the underspecified sense varies greatly, from the very infrequent for English in Animal/Meat to the two Danish datasets where the underspecified sense evens with the metonymic one. However, the Danish examples have mostly three annotators, and chance disagreement is the highest for this language in this setup, i.e., the chance for an underspecified sense in Danish to be assigned by our backoff strategy is the highest.

Columns V and B show respectively whether the underspecified senses are a result of majority voting or backoff. In contrast to volunteers, turkers disprefer the underspecified option and most of the English underspecified senses are assigned by backoff. However, it cannot be argued that turkers have overused clicking on the first option (a common spamming behavior) because we can see that two of the English dot types (eng:artinfo, eng:procres) have majority of metonymic senses, which are always second in the scheme (cf. Fig. 1). Looking at the amount of underspecified senses that have been obtained by majority voting for Danish and Spanish, we suggest that the level of abstraction required by this annotation is too high for turkers to perform at a level compara-

[2]We have made the data freely available at http://metashare.cst.dk/repository/search/?q=regular+polysemy

ble to that of our volunteer annotators.

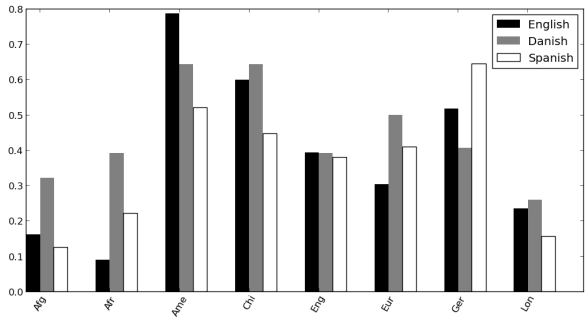

Figure 2: Proportion of non-literality in location names across languages

Figure 2 shows the proportion of non-literal (metonymic+underspecified) examples for the Location/Organization words that are common across languages. We can see that individual words show sense skewdness. This skewdness is a consequence of the kind of text in the corpus: e.g. *America* has a high proportion of non-literal senses in the ANC, where it usually means *"the population or government of the US"*. Similarly, it is literal less than 50% of the times for the other two languages. In contrast, *Afghanistan* is most often used in its literal location sense.

6 Assigning senses with MACE

Besides using majority voting with backoff , we use MACE (Hovy et al., 2013) to obtain the sense tag for each example.

Dot type	L	M	U	D	I
eng:animeat	340	146	14	.048	3
eng:artinfo	170	180	150	.296	46
eng:contcont	295	176	28	.174	0
eng:locorg	291	193	16	.084	3
eng:procres	155	210	134	.272	33
da:contcont	223	134	143	.242	79
da:locorg	251	144	105	.206	53
spa:contcont	270	155	75	.074	56
spa:locorg	302	146	52	.038	40

Table 4: Sense distributions calculated with MACE, plus **D**ifference and **I**ntersection of underspecified senses between methods

MACE is an unsupervised system that uses Expectation-Maximization (EM) to estimate the competence of annotators and recover the most likely answer. MACE is designed as a Bayesian network that treats the "correct" labels as latent variables. This EM method can also be understood as a clustering that assigns the value of the closest calculated latent variable (the sense tag) to each data point (the distribution of annotations).

Datasets that show less variation between senses calculated using majority voting and using MACE will be more reliable. Along the sense distribution in the first three columns, Table 4 provides the proportion of the senses that is different between majority voting and MACE (D), and the size of the intersection (I) of the set of underspecified examples by voting and by MACE, namely the overlap of the U columns of Tables 3 and 4.

Table 4 shows a smoother distribution of senses than Table 3, as majority classes are downweighted by MACE. It takes very different decisions than majority voting for the two English datasets with lowest agreement (eng:artinfo, eng:procres) and for the Danish datasets, which have the fewest annotators. For these cases, the diferences oscillate between 0.206 and 0.296.

Although MACE increases the frequency of the underspecified senses for all datasets but one (eng:locorg), the underspecified examples in Table 3 are not subsumed by the MACE results. The values in the I column show that none of the underspecified senses of eng:contcont receive the underspecified sense by MACE. All of these examples, however, were resolved by backoff, as well as most of the other underspecified cases in the other English datasets. In contrast to the voting method, MACE does not operate with any theoretical assumption about the relation between the three senses and treats them independently when assigning the most likely sense tag to each distribution of annotations.

7 Comparison between methods

The voting system and MACE provide different sense tags. The following examples (three from eng:contcont and four from eng:locorg) show disagreement between the sense tag assigned by voting and by MACE:

d) *To ship a **crate** of lettuce across the country, a trucker needed permission from a federal regulatory agency.*

e) *Controls were sent a package containing stool collection **vials** and instructions for collection and mailing of samples.*

f) *In fact, it was the social committee, and our chief responsibilities were to arrange for bands and **kegs** of beer .*

g) *The most unpopular PM in **Canada**'s modern history, he introduced the Goods and Services Tax , a VAT-like national sales tax.*

h) *This is **Boston**'s commercial and financial heart , but it s far from being an homogeneous district [...]*

i) ***California** has the highest number of people in poverty in the nation — 6.4 million, including nearly one in five children.*

j) *Under the Emperor Qianlong (Chien Lung), Kangxi's grandson, conflict arose between **Europe**'s empires and the Middle Kingdom.*

All of the previous examples were tagged as underspecified by either the voting system or MACE, but not by both. Table 5 breaks down the five annotations that each example received by turkers in literal, metonymic and underspecified. The last two columns show the sense tag provided by voting or MACE.

Example	L	M	U	VOTING	MACE
d)	2	2	1	U	L
e)	3	1	1	L	U
f)	1	2	2	M	U
g)	2	2	1	U	M
h)	2	2	1	U	M
i)	3	0	2	L	U
j)	1	2	2	M	U

Table 5: Annotation summary and sense tags for the examples in this section

Just by looking at the table it is not immediate which method is preferable to assign sense tags in cases that are not clear-cut. In the case of i), we consider the underspecified sense more adequate than the literal one obtained by voting, just like we are also more prone to prefer the underspecified meaning in f), which has been assigned by MACE. In the case of h), we consider that the strictly metonymic sense assigned by MACE does not capture both the organization- ("commercial and financial") and location-related ("district") aspects of the meaning, and we would prefer the underspecifed reading. However, MACE can also overgenerate the underspecified sense, as the vials mentioned in example e) are empty and have no content yet, thereby being literal containers and not their content.

Examples d), g) and h) have the same distribution of annotations—namely 2 literal, 2 metonymic and 1 underspecified—but d) has received the literal sense from MACE, whereas the other two are metonymic. This difference is a result of having trained MACE independently for each dataset. The three examples receive the underspecified sense from the voting scheme, since neither the literal or metonymic sense is more present in the annotations.

On the other hand, e) and i) are skewed towards literality and receive the literal sense by plurality without having to resort to any backoff, but they are marked as underspecified by MACE.

8 Conclusions

We have described the annotation process of a regular-polysemy corpus in English, Danish and Spanish which deals with five different dot types. After annotating the examples for their literal, metonymic or underspecified reading, we have determined that this scheme can provide reliable (α over 0.60) annotations for one dot type. Not all the dot types are equally easy to annotate. The main source of variation in agreement, and thus annotation reliability, is the dot type itself. While eng:animeat and eng:locorg appear the easiest, eng:artinfo and eng:procres obtain very low α scores.

9 Further work

After collecting annotated data, the natural next step is to attempt class-based word-sense disambiguation (WSD) to predict the senses in Tables 3 and 4 using a state-of-the-art system like Nastase et al. (2012). We will consider a sense-assignment method (voting or MACE) as more appropriate if it provides the sense tags that are easiest to learn by our WSD system.

However, learnability is only one possible parameter for quality, and we also want to develop an expert-annotated gold standard to compare our data against. We also consider the possibility of developing a sense-assignment method that relies both on the theoretical assumption behind the voting scheme and the latent-variable approach used by MACE.

Acknowledgments

The research leading to these results has been funded by the European Commission's 7th Framework Program under grant agreement 238405 (CLARA).

References

Mette Skovgaard Andersen, Helle Asmussen, and Jørg Asmussen. 2002. The project of korpus 2000 going public. In *The Tenth EURALEX International Congress: EURALEX 2002*.

J. D. Apresjan. 1974. Regular polysemy. *Linguistics*.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Ted Briscoe, Ann Copestake, and Alex Lascarides. 1995. Blocking. In *Computational Lexical Semantics*. Citeseer.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2525–2529.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of NAACL-HLT 2013*.

N. Ide and C. Macleod. 2001. The american national corpus: A standardized resource of american english. In *Proceedings of Corpus Linguistics 2001*, pages 274–280. Citeseer.

J. B. Johannessen, K. Haagen, K. Haaland, A. B. Jónsdottir, A. Nøklestad, D. Kokkinakis, P. Meurer, E. Bick, and D. Haltrup. 2005. Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1):91.

M. Lapata and A. Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

K. Markert and M. Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. In *Proc. of LREC*. Citeseer.

K. Markert and M. Nissim. 2009. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138.

Vivi Nastase, Alex Judea, Katja Markert, and Michael Strube. 2012. Local and global context for supervised and unsupervised metonymy resolution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 183–193. Association for Computational Linguistics.

Malvina Nissim and Katja Markert. 2005. Learning to buy a renault and talk to bmw: A supervised approach to conventional metonymy. In *Proceedings of the 6th International Workshop on Computational Semantics, Tilburg*.

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

B. S. Pedersen, S. Nimb, J. Asmussen, N. H. Sørensen, L. Trap-Jensen, and H. Lorentzen. 2009. Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Thierry Poibeau. 2006. Dealing with metonymic readings of named entities. *arXiv preprint cs/0607052*.

J. Pustejovsky, A. Rumshisky, J. Moszkowicz, and O. Batiukova. 2009. Glml: Annotating argument selection and coercion. In *IWCS-8: Eighth International Conference on Computational Semantics*.

J. Pustejovsky. 1995. The generative lexicon: a theory of computational lexical semantics.

A. Rumshisky, VA Grinberg, and J. Pustejovsky. 2007. Detecting selectional behavior of complex types in text. In *Fourth International Workshop on Generative Approaches to the Lexicon, Paris, France*. Citeseer.

Ekaterina Shutova. 2009. Sense-based interpretation of logical metonymy using a statistical method. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 1–9. Association for Computational Linguistics.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Jorge Vivaldi. 2009. Corpus and exploitation tool: Iulact and bwananet. In *I International Conference on Corpus Linguistics (CICL 2009), A survey on corpus-based research, Universidad de Murcia*, pages 224–239.