Annotation of regular polysemy

An empirical assessment of the underspecified sense

Author: Héctor Martínez Alonso
Supervisors:
Prof. Bolette Sandford Pedersen (Københavns Universitet)
Prof. Núria Bel Rafecas (Universitat Pompeu Fabra)
Submitted: September 30^{th}, 2013

You are Lover of word you write book one day
— anonymous fortune cookie, 2007

Acknowledgements

To my supervisors, for their high agreement.

To my grand/step/parents and the rest of my modestly-sized family, ever-willing to be more supportive than simply baffled.

To my colleagues at the EMNLP@CPH research team, for all their help. The name is lame but you guys are pretty cool.

To many of my friends, colleagues, relatives and casual acquaintances, for accepting being volunteered into annotating my data:

Allan, Ida and Morten from Infomedia, Adriana Fasanella, Albert Jiménez, Alberto García, Àlex Kippelboy, Amanda Miller, Anders Johannsen, Anna Braasch, Anna Gibert, Anna Lara, Anna Raven, Bea Muñoz, Bente Maegaard, Bjørn Nicola Wesel-Tolvig, Bolette Sandford Pedersen, Brunnihild Beast, Cameron Smith, Carla Parra, Carlos Rubio, Claire Joyce, Claus Povlsen, Cristina Nicolàs, Cristina Sánchez Marco, Daniel Julivert, David Morris, Dorte Haltrup Hansen, Eduardo José Paco Mateo, Elena Ballesteros, Elias Gallardo, Elias Portela, Elisabet Vila, Estanis Solsona, Ester Arias, Eva Bosch Roura, Fabio Teixidó, Francesc Torres, Glòria de Valdivia Pujol, Idoia Irazuzta, Inés Plaza, Irene Peralta, Isabel Gamez, Isabel Sucunza, Isabel Sucunza, Jakob Elming, Jaume Cladera, Jesper Kruse, Jimmi Nielsen, Joan Quílez, Joaquín Báñez, Jon Miller, Jordi Allué, Jordi Fortuny, Jordi Marin, Lars-Jakob Harding Kællerød, Lauren Romeo, Leah González, Lene Offersgaard, Line Burholt, Lorena Ronquillo, Magdalena Gómez, Marc Kellaway, Maria Antònia Servera Barceló, Maribel Marín, Melanie Wasserman, Miguel Ángel Fuentes, Miquel Adell, Mireia Giménez, Mogens Daugbjerg Laustsen , Mónica Pérez, Montse Alonso, Morten Bech Sørensen, Morten Mechlenborg Nørulf, Nacho Pintos, Nereida Carillo, Nereya Otieno, Núria Galera, Núria Luna, Ole Bremer Pedersen, Pau Clemente Grabalosa, Paw Skjoldan, Paw Sort Jensen, Pedro Gómez, Pedro Pacheco, Pedro Patiño, Peter Juul Nielsen, Puri Alonso, Randi Skovbjerg Sørensen, Raul Guerrero, Raul Morais, René Kruse, Samuel Riberio, Sandra Arnáiz, Sandra Novillo, Sigrid Klerke, Sune Sønderberg Mortensen, Susana Pagán, Susanne Kemp, Sussi Olsen, Tanya Karoli Christensen, Thomas Johannsen, Toni Mata, Uffe Paulsen, Veronica Casas, Verónica García, Victor Pascual, and Woo Jin Ko.

And to the creators and curators of the twenty-odd language resources and tools I used to carry out the experiments in this dissertation, for making their work freely available.

Funding

This research has received support from the EU 7^{th} Framework Program under a Marie Curie ITN, project CLARA.

Abstract (English)

Words that belong to a semantic type, like LOCATION, can metonymically behave as a member of another semantic type, like ORGANIZATION. This phenomenon is known as *regular polysemy*.

In Pustejovsky's (1995) Generative Lexicon, some cases of regular polysemy are grouped in a complex semantic class called a *dot type*. For instance, the sense alternation mentioned above is the LOCATION•ORGANIZATION dot type. Other dot types are for instance ANIMAL•MEAT or CONTAINER•CONTENT.

We refer to the usages of dot-type words that are potentially both metonymic and literal as *underspecified*. Regular polysemy has received a lot of attention from the theory of lexical semantics and from computational linguistics. However, there is no consensus on how to represent the sense of underspecified examples at the token level, namely when annotating or disambiguating senses of dot types.

This leads us to the main research question of the dissertation: Does sense underspecification justify incorporating a third sense into our sense inventories when dealing with dot types at the token level, thereby treating the underspecified sense as independent from the literal and metonymic?

We have conducted an analysis in English, Danish and Spanish on the possibility to annotate underspecified senses by humans. If humans cannot consistently annotate the underspecified sense, its applicability to NLP tasks is to be called into question.

Later on, we have tried to replicate the human judgments by means of unsupervised and semisupervised sense prediction. Achieving an NLP method that can reproduce the human judgments for the underspecified sense would be sufficient to postulate the inclusion of the underspecified in our sense inventories.

The human annotation task has yielded results that indicate that the kind of annotator (volunteer vs. crowdsourced from Amazon Mechanical Turk) is a decisive factor in the recognizability of the underspecified sense. This sense distinction is too nuanced to be recognized using crowdsourced annotations.

The automatic sense-prediction systems have been unable to find empiric evidence for the underspecified sense, even though the semisupervised system recognizes the literal and metonymic senses with good performance.

In this light, we propose an alternative representation for the sense alternation of dot-type words where literal and metonymic are poles in a continuum, instead of discrete categories.

Abstract (Danish)

Ord som hører til en semantisk klasse som fx STED kan vise en metonymisk betydning som fx VIRKSOMHED. Dette fænomen kaldes for *systematisk polysemi*.

I Pustejovskys (1995) Generative Lexicon forefindes særlige kombinationer af bogstavelig og metonymisk betydning. De samles i en kompleks semantisk klasse eller *dot-type* som for eksempel STED•VIRKSOMHED, DYR•KØD eller BEHOLDER•INDHOLD.

Hvis en anvendelse af et dot-type ord er både bogstavelig og metonymisk, kaldes den for *uspecificeret*. Systematisk polysemi har fået meget opmærksomhed fra teori om leksikal semantik og fra datalingvistik, men der er ikke konsensus om, hvordan den uspecificerede betydning repræsenteres på token-niveau, dvs når man annoterer individuelle betydninger eller udfører *word-sense disambiguation*.

Dette fører til vores hovedforskningsspørgsmål: Skal man behandle den uspecificerede betydning som en enestående betydning, uafhængig af den bogstavelige og den metonymiske betydning, når man arbejder med dot-type ord på token-niveau?

Vi har udført et studie med engelske, danske og spanske eksempler for at forske i den menneskelige evne til at annotere den uspecificerede betydning. Hvis mennesker ikke kan annotere den på en konsistent måde, bør dens anvendelsesmuligheder for datalingvistiske systemer gentænkes.

Senere har vi prøvet forskellige forsøg for at replikere annotationerne med superviseret og usuperviseret maskinlæring. Finder vi et datalingvistisk system, som kan forudse de menneskelige betydningsbedømmelser om uspecificering, har vi tilstrækkeligt bevis for at inkludere den uspecificerede betydning i vores betydningsbeholder.

Annotationsopgaven med mennesker har givet et resultat, som påpeger, at den slags annoter (frivillig eller crowsourced fra Amazon Mechanical Turk) er en afgørende faktor for at kunne genkende de udspecificerede eksempler. Denne betydningsforskel er alt for nuanceret til at kunne genkendes med crowdsourcede annoteringer.

Det automatiske system for betydningsforudsigelse kan ikke finde empirisk bevis for en uspecificeret betydning, selv om systemet præsterer udover det tilstrækkelige for den bogstavelige og den metonymiske betydning.

På denne baggrund foreslår vi en alternativ repræsentation for dot-type ord, hvor bogstavelighed og metonymi er poler i en gradient af kontinuerlige værdier, i stedet for diskrete klasser.

Abstract (Spanish)

Las palabras de una clase semántica como LUGAR pueden comportarse metonímicamente como miembros de otra clase semántica, como ORGANIZACIÓN. Este fenómeno se denomina *polisemia regular*.

En el Generative Lexicon de Pustejovsky (1995), algunos casos de polisemia regular se encuentran agrupados en una clase semántica compleja llamada *dot type*. Por ejemplo, la alternación de sentidos anterior es el dot type LUGAR•ORGANIZACIÓN. Otros ejemplos de dot type son ANIMAL•CARNE or CONTENEDOR•CONTENIDO.

Llamamos *subespecificados* a los usos de palabras pertenecientes a un dot type que son potentialmente literales y metónimicos. La polisemia regular ha recibido mucha atención desde la teoría en semántica léxica y desde la lingüística computacional. Sin embargo, no existe un consenso sobre cómo representar el sentido de los ejemplos subespecificados al nivel de token, es decir, cuando se anotan o disambiguan sentidos de palabras de dot types.

Esto nos lleva a la principal pregunta de esta tesis: ¿Justifica la subespecificación la incorporación de un tercer sentido a nuestros inventarios de sentidos cuando tratamos con dot types a nivel de token, tratando de este modo el el sentido subespecificado como independiente de los sentidos metónimico y literal?

Hemos realizado un análisi en inglés, danés y español sobre la posibilidad de anotar sentidos subespecificados usando informantes. Si los humanos no pueden annotar el sentido subespecificado de forma consistente, la aplicabilidad del mismo en tareas computacionales ha de ser puesta en tela de juicio.

Posteriormente hemos tratado de replicar los juicios humanos usando aprendizaje automático. Obtener un método computacional que reproduzca los juicios humanos para el sentido subespecificado sería suficiente para incluirlo en los inventarios de sentidos para las tareas de anotación.

La anotación humana ha producido resultados que indican que el tipo de anotador (voluntario o *crowdsourced* mediante Amazon Mechanical Turk) es un factor decisivo a la hora de reconocer el sentido subespecificado. Esta diferenciación de sentidos requiere demasiados matices de interpretación como para poder ser anotada usando Mechanical Turk.

Los sistemas de predicción automática de sentidos han sido incapaces de identificar evidencia empírica suficiente para el sentido subespecificado, a pesar de que la tarea de reconocimiento semisupervisado reconoce los sentidos literal y metonímico de forma satisfactoria.

Finalmente, propones una representación alternativa para la representación de sentidos de las palabras de dot types en la que literal y metónimico son polos en un contínuo en lugar de categorías discretas.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Generative Lexicon (GL) offers itself as an alternative to the traditional, sense-enumeration based understanding of word senses, by both postulating underspecified lexical entries and the possibility of underspecified predications. Some computational lexicons have been developed using the Generative Lexicon as a theoretical framework and incorporate the dot type within their hierarchy of semantic classes.

While the analytical usefulness of postulating the dot type has been proven, there is no larger scale empirical assessment of the statistical significance of metonymic and underspecified predications of dot-type nouns.

There are few approaches from within statistical machine learning (ML) that make use of the GL as, for instance, the listing of target classes for classification experiments in word sense disambiguation. Computational semantics employs mostly clustering and classification, that is, methods that assign an input document a category from a list of mutually exclusive categories. How can we then develop applications following a theory that transcends the sense-enumeration paradigm if we are straitjacketed by our computational modus operandi, bound to provide a discrete output?

While reading the Generative Lexicon (GL), I was intrigued by the parallel notions of dot type and regular polysemy. Was the dot type an adequate means to describe regular polysemy at the token-level for natural language processing (NLP) tasks? In other words, was it useful or even possible to annotate the underspecified senses in cases of regular polysemy using GL theoretical objects? And to which extent were the concepts offered in the GL implementable in current methods within computational semantics?

From these considerations arose the general research question in this dissertation. To which extent are underspecified predications of dot-type words an actual class-wise distributional phenomenon that NLP systems need to account for, and not an unnecessary theoretical artifact resulting from the attempt to develop a theoretical framework to encompass both predicate logic and lexical semantics?

The main goal of this thesis is the empirical assessment of the relevance of the dot type for token-level sense assignment. The lack of a sense-annotated corpus aiming at capturing the selectional behavior of dot types—and particularly, of underspecified predications—called for the elaboration of a corpus with sense annotations for dot type words.

The dot type is a theoretical object of the GL, and in order to study the viability of its usage for sense annotation, we needed to choose specific instances of dot types. We have chosen to conduct the study on a series of dot types for Danish, English and Spanish. We have chosen five dot types to study for English, and picked two of them for a comparative in Danish and Spanish.

We developed a sense-annotated corpus for a series of dot types in Danish, English and Spanish. This sense-annotated corpus served a dual purpose; first, it allowed an empirical study on how different kinds of annotators (experts, turkers, and volunteers) are predisposed to recognizing the underspecified sense; second, it served as gold standard for the following NLP experiments.

Collection of data is only but one step in the development of a gold standard. A great deal of the effort documented in this dissertation went to providing strategies to capture underspecified predications from the (again, discrete and categorical) judgments of the informants. One of the strategies was a voting method with a theoretically motivated backoff strategy, and the other an unsupervised method to automatically weigh annotator judgments. Both sense assignment methods were compared against the expert annotations to determine which one yielded more reliable sense assignments.

Once the gold standard data had been assembled, it was available it for machine learning tasks. The general goal of computational semantics is to automatically emulate the human informants' meaning judgments, and the experiments try to capture different sides of the regular polysemy and underspecification phenomena.

Since a way of proving the existence of a pattern is demonstrating its learnability by automatic means, this thesis experimented with systems that aimed at the disambiguation of senses within a list of dot types.

A first unsupervised system used word-sense induction to account for the distributional evidence behind the different senses in a dot type. Its goal is to assess how much simple distributional information there is to identify the senses of dot type words, especially whether there is empirical evidence to claim an underspecified sense.

Second I used a (semi-)supervised approach to classify individual predicates into literal, metonymic and underspecified. This word sense disambiguation experiment was aimed to assess the identifiability of underspecified predications.

Two experiments complement this study. The first one tries to represent the notion of a continuum between the literal and the metonymic as a numeric value that can be calculated. Its goal is to assess the theory that describes literality and metonymicity as a continuum and defends the gradability of the metonymy readings as non-categoric, non-discrete, non-mutually exclusive.

The second one is a similar regression experiment to try to predict the observed agreement of examples. The difficulty of an example in terms of vocabulary or syntactic complexity can also hinder the annotator's agreement, thus affecting the distribution of underspecified predications. Low agreement does not mean useless information, but indicates difficulty and is potentially correlated to underspecified sense. This experiment can thus account for the linguistic traits that make an example more or less difficult to annotate.

Chapter 2 defines the concepts and terms that are necessary for this dissertation, namely a definition of regular polysemy, an overview of the pertinent theoretical objects of the Generative Lexicon, and redefines the research question in a more formal manner. Chapter 3 covers the related work in the

representation of metonymy in lexical knowledge bases, sense-annotation tasks for figurative meaning, and type- and token-based modeling of regular polysemy. Chapter 4 describes the annotation task to obtain human judgments on regular polysemy; this chapter also compares the biases of expert, turker and volunteer annotators for the underspecified sense, and it describes the implementation of two sense assignment methods to obtain a final sense tags from the set of judgments given by the annotators. Chapter 5 describes the word-sense induction method that tries to capture the senses of dot types by unsupervised means. Chapter 6 describes the features we use to characterize the examples for the (semi-)supervised experiments of the following chapters. Chapter 7 describes the word-sense disambiguation system that aims to identify the literal, metonymic and underspecified senses. Chapter 8 describes an experiment to assess the robustness of a continuous representation of the token-wise sense of dot-type words. Chapter 9 describes an experiment to predict the agreement of sense-annotated examples. Finally, Chapter 10 summarizes the conclusions from the experiment chapters, lists the additional contributions of the dissertation, and outlines the future work.

Chapter 2

Theoretical framework

In this chapter we define the key concepts—and the corresponding terms— that we need in order to conduct the empirical study outlined in the Introduction. Section 2.1 defines the notion of polysemy, which is expanded in Section 2.2 with the notion of regular polysemy. In Section 2.3 we describe a non-discrete understanding of the gradation between literal and figurative sense. In Section 2.4 we define the concept of underspecification with regards to regular polysemy. Section 2.5 provides an overview of the Generative Lexicon. In Section 2.6 we describe how we choose particular cases of metonymy to conduct our study, and finally in Section 2.7 we redefine the main research question of this dissertation, broken down into three hypotheses.

2.1 Polysemy

Words can have more than one meaning. There is not, however, a single cause for the appearance of multiple meanings. In some cases, a word—and by *word* we mean a sequence of letters bound by separators—can present meanings that are unrelated to each other, like the "sport competition" and the "little flammable stick" senses of the noun *match*, the former being of Germanic origin and the later being of Romance origin:

match *noun*
1: *a person or thing equal or similar to another*
2: *a pair suitably associated*
3: *a contest between two or more parties, "a golf match", "a soccer match", "a shouting match"*
...

match *noun*
1: *a chemically prepared wick or cord formerly used in firing firearms or powder*
...

We refer to words that show unrelated meanings but share surface form as *homonyms*, as the common understanding of these phenomenon is that there are two or more concepts that share surface form, even though they have no

morphosyntactic, denotational or etymological relation. In these cases it is customary to give each meaning its own entry in a dictionary.

Merriam-Webster has indeed two separate entries for the noun *match*, one for each of the homonymic meanings—there is also an entry for the verb *match*, but we are disregarding it for the sake of the argument. These entries are further broken down to define more fine-grained senses. Let us examine the entry for *"match (flammable stick)"*:

match *noun*
1: *a chemically prepared wick or cord formerly used in firing firearms or powder*
2: *a short slender piece of flammable material (as wood) tipped with a combustible mixture that bursts into flame when slightly heated through friction (as by being scratched against a rough surface)*

This entry provides two senses. The second one is the most conventional nowadays, whereas the first one is the older, original sense. The two senses are etymologically related and thus, we consider this a case of *polysemy*, and not homonymy.

Polysemy is different to homonymy in that the senses of polysemous words share *some kind* of relation between them. Intuitively, we can see that the "wick used in firing firearms" and the "piece of flammable material" senses of *match* are closer to each other than to the "sports competition" sense.

This later sense also coexists in a polysemous grouping with other related senses of *match* like "an exact counterpart" or "a marriage union". We see thus that polysemy and homonymy can and do coexist, which obscures the application of this distinction, because for the same word we find two clusters of internally related senses; one for senses that have to do with lighting fire, and one for senses that have to do with pairings of comparable members.

If the criteria to differentiate between the two clusters are often etymological, this distinction becomes more of a philological concern than an actual issue within computational linguistics, where—in its most basic conceptualization—each word has a list of potential, mutually exclusive, unrelated senses.

Let us compare the senses in the previous entries with the first four senses in the WordNet entry for *match*:

match *noun*

1. noun.artifact: match, lucifer, friction match (lighter consisting of a thin piece of wood or cardboard tipped with combustible chemical; ignites with friction)
2. noun.event: match (a formal contest in which two or more persons or teams compete)
3. noun.artifact: match (a burning piece of wood or cardboard)
4. noun.artifact: match, mate (an exact duplicate)

We see that these senses are provided as a list, even though they could be further structured by assigning them to the two clusters mentioned previously, issuing groups {1,3} and {2,4}. Lexical knowledge bases like WordNet have been criticized by authors like Pustejovsky and Boguraev (1996) or Kilgarriff (1997) because of their way of portraying and listing senses as an enumeration of unrelated definitions (cf. Section 2.5.4).

However, we know that homonymy is arbitrary, whereas the senses of polysemous words are somehow related. If these senses are related, it is because they are the result of *meaning shifts* that modify and nuance the meaning of words, thereby changing the meaning of words on the fly as required to fit the needs for expression. Some of these meaning shifts become conventionalized and added to the fix repertoire of possible expectable senses a word can manifest.

The concern in this dissertation is not to explore the mechanisms by which new figurative senses are conventionalized, but rather, to explore how words of certain classes like animals or places show a related, secondary sense that is consistent throughout the whole class. In this way, all the names of inhabited places can also be used to refer to their population, and most animal names can be used to refer to their meat. This particular kind of polysemy is called *regular polysemy*

Apresjan (1974) offers a formalization of the intuition of coherence between senses we have mentioned above:

> The meanings a_i and a_j of the word A are called similar if there exists levels of semantic analysis on which their dictionary definitions [...] or associative features have a non-trivial common part.

Still, he also concedes that the difference between homonymy and polysemy is not always immediate but suggests a gradation from homonymy to traditional ("irregular") polysemy to regular polysemy in terms of how coherent the senses of the words are:

> Polysemy and homonymy are relative concepts [...] On the other hand, we can speak of different types of polysemy from the point of view of their relative remoteness from homonymy.

Apresjan conceptualizes the three phenomena of homonymy, irregular and regular polysemy as being closer or further apart from each other if the senses of the words are respectively closer or further apart from each other:

> Closest to homonymy are some types of metaphorically motivated polysemy; in some dictionaries such polysemy is typically treated as homonymy. [...] Metonymically and functionally motivated polysemy is, generally speaking, removed a step further from metonymy.

Figure 2.1 represents this statement graphically. Homonymy is the case where the senses are in general less related to each other, because they are arbitrarily grouped under the same surface form—i.e. the same word. Irregular polysemy, often based on metaphor (cf. Section 2.3) relates senses in a more coherent manner than homonymy, and regular polysemy shows the highest internal coherence of the three because its senses are obtained from metonymic processes more often than not.

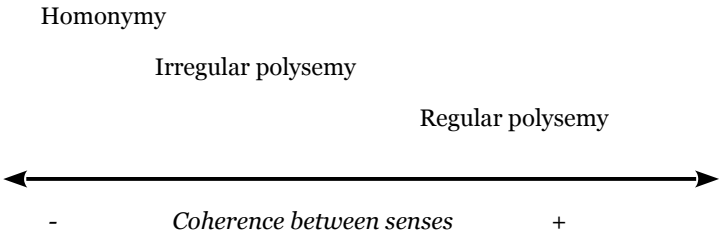

Figure 2.1: Graphic representation of homonymy and the two polysemies along an idealized axis of semantic coherence between senses

The next sections provide a formalized definition of the two different kinds of polysemy, focusing on regular polysemy.

2.2 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):

(2.1) 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 to England as a political entity (ORGANIZATION). The third case refers to both the English territory and the English people. The ability of certain words to switch between semantic types in a predictable manner is named by different authors as *logical metonymy* (Lapata and Lascarides, 2003), *sense extension* (Copestake and Briscoe, 1995), *transfer of meaning* (Nunberg, 1995), *logical* or *complementary polysemy* (Pustejovsky, 1995), *systematic polysemy* (Blutner, 1998) or regular polysemy.

Along with all these terminological variants comes a plethora of definitions, each highlighting a particular aspect of this phenomenon of semantic variation. Let us compare the definition given by Apresjan with Pustejovsky's. Apresjan (1974) offers the two parallel definitions. One of them notes that

> Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B

> with the meanings b_i and b_j, which are semantically distinguished from each other in the same way as a_i and a_j and a_i and b_j are nonsynonymous.

In the second definition, he adds that

> Polysemy is called irregular if the semantic distinction between a_i and a_j is not exemplified in any other word of the given language.

In other words, according to Apresjan the polysemy encompassing the "English territory" (a_i) and the "English government" (a_j) senses of *England* is regular because we can find a similar pattern for Portugal (b_i ="the Portuguese territory";b_j ="the Portuguese government") and for any other state. In order to operationalize this definition we need to understand it as a collective phenomenon that happens at the semantic type or class level, and not only to individual words. Pustejovsky (1995) provides a different wording:

> I will define logical polysemy as a complementary ambiguity where there is no change of lexical category, and the multiple senses of the word have overlapping, dependent or shared meanings.

Pustejovsky's definition is word-internal and focuses on the relations between the senses and the possibility of overlap between them in the cases of regular polysemy, whereas Apresjan focuses on the consistency of alternation of senses across all the other members of the same semantic class.

Apresjan acknowledges that polysemy needs to be present in a plural number of words to be regular, that is, words A and B in the previous example form a class by virtue of their first sense. Polysemy is thus considered regular if it is extensive to the semantic type of the polysemous word and not only to the word itself.

Moreover, Apresjan isolates regular polysemy from the other polysemies by calling the later irregular. He later adds that "regularity is a distinctive feature of metonymical transfers, irregular polysemy is more typical of metaphorical transfers", in this way defining both how to identify regular polysemy—by regularity in sense alternations across words—and what processes cause it—mostly metonymy but also metaphor (cf. Section 2.3).

After these considerations we define *regular polysemy* as a phenomenon whereby the word that belongs to a semantic type can predictably act as members of another semantic type. Since there is a change of semantic type (and words can stop meaning what they originally mean), this semantic shift is a case of figurative language. The definition of regular polysemy we use in this dissertation rests on the notions of metaphor and especially metonymy; both are covered in Section 2.3.

2.3 Figurative language

Traditional rhetoric describes phenomena that allow using an expression to mean something of a distinct kind, licensing *figurative* uses of language. Nunberg (1995) calls these phenomena *transfers of meaning*, because some element of the meaning of the original expression is transferred to the intended sense. Beyond expressions based on degree like understatement or hyperbole, there are two

main methods to extend the meaning of a word beyond its literal sense, namely metaphor and metonymy.

Metaphor is commonly used as the default form of figurative language, and there are definitions that incorporate metonymy as a kind of metaphor. Our definition sees metaphor and metonymy as two different daughter categories to figurative language that do not subsume each other.

Metaphor is a method to extend the sense of words by exploiting **analogy** or similarity between concepts. *Metonymy* is a method to extend the sense of words exploiting **contiguity**. Dirven (2002) places the origin of this differentiation in Roman Jacobson's work (Jakobson, 1971). Historically, he claims, linguistic characterization had focused on metaphor and "metonymy tended to be very much neglected" until Lakoff and Johnson (1980) published their canonical work *Metaphors We Live By*, that in spite of its focus on metaphor, also provides a definition of metonymy:

> In these cases, as in the other cases of metonymy, one entity is being used to refer to another. Metaphor and metonymy are different kinds of processes. Metaphor is principally a way of conceiving one thing in terms of another, and its primary function is understanding. Metonymy, on the other hand, has primarily a referential function, that is, it allows us to use one entity to stand for another

Lakoff and Johnson understand metaphor in terms of domain mapping, in an attempt to quantify the difference between analogy and contiguity. If a domain is a realm of experience (TEMPERATURE, TIME, SPACE, etc.), we can define analogy as the ability to transfer certain properties of one domain to another. For instance, *heat* is a term of the TEMPERATURE domain which can be called *"high temperature"*, *high* being a concept from the SPACE domain which is transferred following the notion of MUCH IS HIGH.

Metonymy however is conscribed within one domain—that of the base semantic type of the metonymic word—, which is what determines the contiguity mentioned above. Metonymic senses refer to something **physically adjacent**, like a part of a larger whole, or to something **temporally adjacent**, like a consequence or a product. Some well-known examples of metonymic sense alternations are:

(2.2) a) PART FOR WHOLE: Lots of *hands* were needed to fix the fence.

b) CONTAINER FOR CONTENT: He drank a whole *glass*.

c) LOCATION FOR ORGANIZATION: *France* elects a new president.

d) PROPERTY FOR SUBJECT OF PROPERTY: The *authorities* arrived quickly.

e) PRODUCER FOR PRODUCT: I drive a *Honda*.

When reading example a), a part-for-whole metonymy (also known as *synecdoche*), we know that "hands" stands for "working people". In the rest of the examples in (2.2), the word in italics also stands for something that is not exactly its original sense: glasses are solid and cannot be drunk, France is a place and has no will of its own, authority is abstract and does not move, etc. Still, these are examples of very usual metonymies that do not convey a strongly poetic or non-conventional sense.

Even though we refer to both metaphor and metonymy as figures of speech, we support the intuition that there is something in metonymy that feels closer to literality, that is, closer to conventional meaning (cf. 2.3.2). This intuition is consistent with the idea of contiguity in metonymy as opposed to analogy in metaphor, and with the idea that metaphors are mappings across two domains while metonymy is in-domain.

Gibbs (1984) proposes that the literal and figurative meanings are placed at the poles of a single continuum where intermediate senses are spread. Dirven (2002) offers a taxonomy of the degrees of non-literality from literal (left) to completely figurative (right). Figure 2.2 shows a summary of the schema provided in Dirven (2002). This understanding is also found in Cruse (Cruse, 1986, p. 71) as *sense spectrum*.

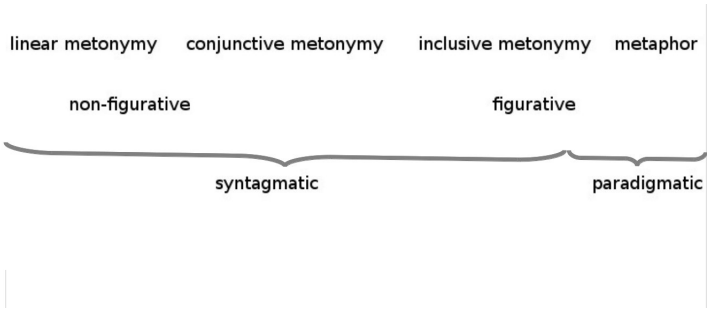

Figure 2.2: The figurative-literal continuum according to Dirven (2002)

Let us focus on Dirven's gradation of figurativeness of metonymies by looking at his examples:

(2.3) a) Different parts of the *country* do not mean the same.
b) *Tea* was a large meal for the Wicksteeds.
c) The *Crown* has not withheld its assent.
d) He has a good *head* on him.

Example a) shows what Dirven calls lineal metonymy, which he considers non-figurative and non-polysemous (cf. Section 2.4). If there is a chance for metonymies showing polysemous senses that coexist in a predication, this is known by Dirven as conjunctive metonymy in b) and d), even though he perceives the second example as non-figurative and the third one as figurative. The last example is considered inclusive polysemy and is always figurative and polysemous. Dirven calls *polysemous* a word that has simultaneously literal and metonymic usage. We expand on this notion in Section 2.4.

The axis in Figure 2.2 corresponds to the idea of internal consistency of senses across homonymy and polysemy in Figure 2.1. That is, the more figurative the figurative sense, the longer the distance from the original meaning will be. In a conceptually related work, Hanks (2006) expands the right side of this continuum by postulating that some metaphors are more figurative than others. For instance, he claims that "a desert" in "a desert of railway tracks" is less figurative than "I walked in a desert of barren obsession".

2.3.1 Differentiating metonymy from metaphor

In the previous section we have given definitions of metaphor and metonymy and placed them in a literal-figurative continuum. The differences between metaphor and metonymy indicate how the non-literal senses of each kind are predicated. Warren (2002) provides a list of differences between metaphor and metonymy, from which we highlight three that are relevant for us:

1. Metaphor involves seeing **something in terms of something else**, whereas metonymy does not. Warren mentions:

   > That is to say, metaphor is hypothetical in nature. Life is thought as if it were a journey. Metonymy, on the other hand, is not hypothetical. There is nothing hypothetical about *the kettle* in *the kettle is boiling*, for instance. It is for this reason that I make the point that non-literalness in the case of metonymy is superficial.

   This remark is relevant because it supports the postulation of regularity in Section 2.2, as the normalcy of usage of metonymy aids its regularity. Moreover, this statement is also a support for our understanding of metonymy as a paraphrase in Section 2.3.3.

2. Metaphors can form **themes**. Warren gives an example of a text based around the metaphor MEMBERS OF PARLIAMENT ARE WELL-TRAINED POODLES, where British politicians are described licking the boots of the Prime Minister and being rewarded with a biscuit. She also acknowledged that metonymic patterns can be conventionalized (CONTAINER FOR CONTENT), but claims that these metonymic patterns never give rise to themes like the ones found in her examples. This remark is relevant because it justifies annotating at the sentence level instead of annotation larger discourse segments: if metonymy does not originate themes at the document level, it is safe to conduct our empirical study at the sentence level (cf. 4.2).

3. Metonymies can coexist in **zeugmatic constructions** whereas metaphors cannot. She considers the metonymic example "Caedmon is a poet and difficult to read" a valid sentence, while she rejects the example "?The mouse is a favorite food of cats and a cursor controller" because it incurs in zeugma, namely failing at coordinating this two concurrent senses. This remark is relevant because it allows the zeugma test in the expert annotation scheme described in Section 4.3.

Even though metaphor and metonymy are mechanisms for figurative language, they are qualitatively different, because metaphor is a way to view one entity in terms of another, reasoning by analogy; whereas metonymy is a way to get an entity to stand for another, reasoning by contiguity.

In this dissertation we deal strictly with cases of regular polysemy that are triggered by metonymic processes and not by metaphor (cf. Section 2.6 for our object of study). However, there are cases of very productive metaphors giving rise to polysemous senses that are pervasive at the semantic class level, thus ensuing regular polysemy.

(2.4) a) REPRESENTATION FOR OBJECT: She wore a skirt full of *flowers*.

b) BODY PART FOR PART OF OBJECT: I hit the *foot* of the lamp.

c) ARTIFACT FOR ABSTRACT PROPERTY: Human population is a *timebomb*.

In the first example of 2.4, "flowers" stands for "drawings of flowers". We consider this example to be a metaphor, unlike Markert and Nissim (2002b) who consider it a metonymy. But insofar a visual representation of something is motivated by analogy, it is a metaphor. This polysemy is regular because any physical object could be represented and drawn or embroidered on the skirt, be it birds, stars or hot dogs for that matter; but it is still metaphorical because it is based on analogy and because it cannot be copredicated: "?She wore a skirt full of flowers and they smelled lovely" incurs in zeugma.

Example b) illustrates that the ability of body-part words to be used to delimit parts of objects that are not human bodies is also common. This very usual metaphor has also issued conventionalized metaphoric uses of words like *neck* for the open end of a bottle or *hand* for the dials of a clock.

In example c), *timebomb* stands for something bad that is going to happen at some point in the future. This is an example of regular polysemy because using artifact words to transfer some of the properties of the artifact unto the described phenomena is very productive, and an obvious example of the two-domain approach—one domain being POPULATION and the other one being EXPLOSIVES—described by Lakoff and Johnson (1980). In a corpus study over Danish texts, Nimb and Pedersen (2000) provide an account on similar metaphors with words like *fan*, *bridge* or *springboard*.

It is also worth mentioning how the metaphor in c) could issue a theme as explained at the beginning of this section. Another sentence in the same text could be "the environment is going to suffer from its detonation", thereby following the explosive theme when talking about the detonation of the metaphoric population timebomb.

2.3.2 Literal, abstract and concrete meaning

In the previous section we have mentioned literal meaning, largely taking such a notion for granted. Talking about metaphor assumes there is a literal meaning that gets partially transferred from one domain to the other when using a word in a metaphoric manner. Likewise, metonymic senses are extensions of the literal sense to something that is spatially or temporally contiguous to the physical referent for the original word.

The notion of literal meaning has however its detractors, like Gibbs (1984), who claims that the distinctions between literal and figurative meanings have little psychological validity, or Hanks (2004), who subscribes to an understanding of lexical meaning that denies that words have meaning at all unless they are put into context.

Still, Hanks (2006) concedes that the expression "literal meaning of a word" is useful, provided that not too much theoretical weight is put on it. We incorporate the notion of literal sense because it allows us to determine that one of the senses in a metonymic alternation is basic or fundamental—i.e. literal—, and the other one metonymic.

We consider the *literal* meaning the most conventional and prototypical sense of a word. Literal meaning will be differentiated from *figurative* meaning in this dissertation, and in most cases this figurative meanings will be metonymic. The alternating senses in the cases of regular polysemy displayed in this work will either be called *literal* or *metonymic*.

In order to identify what is the literal sense of a possibly metonymic word, and thus the directionality of the metonymy, Cruse (Cruse, 1986, p. 69) suggests a negation test of the "half" that is predicated to test for directionality. He provides the following examples:

(2.5) a) I'm not interested in the binding, cover, typeface etc.—I'm interested in the novel

b) ?I'm not interested in the plot, characterisation, etc.—I'm interested in the novel

Cruse considers that it is not anomalous to say a), whereas b) holds a semantic dissonance. Cf. Section 2.6 for the complications some metonymic phenomena pose when determining the literal sense.

While there is a correlation between figurative usage of words and the abstractness of the context in which they appear (Turney et al., 2011), figurativeness and abstraction are different notions. It is therefore important to make a distinction between senses being literal or figurative, and senses being concrete or abstract separately.

We define *concreteness* as a trait that all the names for physical objects have in common, that is, physical denotation. Entities like *car*, *sea* or *penguin* are concrete. These entities are also called first-order entities by Lyons (1977).

Furthermore, we pool together Lyons' second-order entities (events like *sunset* and *invasion*) and third-order entities (mental entities like *reason* or *theorem*) into the group of *abstract* nouns. This issues a coarse grouping, as intuitively an *invasion* (second-order) is more concrete than an *idea* (third-order). Hanks (2006) refers very possibly to third-order entities when he claims that abstract nouns are not normally used to make metaphors. But on the other hand, an *invasion* is arguably more abstract than a *sunset* and they are both second-order entities. We will in general stick to this differentiation between concrete and abstract and mention second- or third-order entities when we deem it convenient. For more on literal meaning, its definition and pragmatics, cf. Recanati (1995, 2002).

2.3.3 Metonymy and paraphrase

In this section we describe the relation between metonymy and paraphrase. In a phrasing like the one offered by Lakoff and Johnson (1980):

> Metonymy has primarily a referential function, that is, it allows us to use one entity to stand for another.

Lakoff and Johnson's remark indicates that metonymy is a pragmatic operation. Since metonymy exploits world knowledge of physical or temporal contiguity, it deviates from being a strictly semantic operation like metaphor, where something is explained in terms of something else, in favor of being a pragmatic operation (Hobbs and Martin, 1987; Hobbs et al., 1993).

Stallard (1993) holds that the felicitousness of a metonymy depends essentially on the Gricean Maxim of Quantity (Grice, 1975). Also Pustejovsky (1995) explains the interpretation of regular polysemy through lexical defaults associated with the noun complement in the metonymic phrase. In essence both explanations entail the same, namely that uttering or writing a metonymy implies exploiting world (or lexical) knowledge to use fewer words and make conversation or reading more fluid or expressive.

Paraphrase as such can only be postulated as a relation between two expressions with the same meaning (Recasens and Vila, 2010), but most of the interpretations of metonymies in this dissertation will be a paraphrase where the sense of the metonymic word is expanded to include the information that we understand was held back (cf. Section 3.3).

The systematicity of regular polysemy, which is a class-wise phenomenon, is supported by the conventionality of the Gricean maxim of quantity. The maxim offers a pragmatic motivation for the expectability of statements like "I drank the glass" over "I drank what the glass contained" (or another less awkward wording of a non-metonymic paraphrase), where the sender holds information back that the receiver can resolve.

The pragmatic motivation for using metonymies as paraphrases also justifies the sense-annotation method described in Section 4.3, where the expert annotator paraphrased the possibly metonymic headword to disambiguate and examined how felicitous the statement was. Paraphrasing as a method to identify potentially figurative senses has also been used by Nissim and Markert (2005a).

As a closing terminological remark, and in line with the Gricean Maxim of Quantity, in this dissertation we will use extensively the term *regular polysemy* to mean "metonymy-based regular polysemy". If a distinction is necessary we will refer to the other kind as *metaphor-based regular polysemy*.

2.4 Underspecification

As we have seen in Section 2.3, metonymic senses can be coordinated together with their fundamental literal sense in the same sentence without incurring in zeugma:

(2.6) a) *Lunch* was delicious but took forever.

b) *Shakespeare* has been dead for centuries and people still read him.

In the first example in (2.6), we have "but" coordinating the statements "lunch was delicious"—in which *lunch* means food—and "lunch took forever"—in which *lunch* means the mealtime, an event. In the second example, Shakespeare means "the person William Shakespeare" but also the metonymic sense "the works of William Shakespeare", even though the second clause has a pronoun to stand for Shakespeare.

The possibility to coordinate the two alternating senses is the key linguistic test to differentiate metaphors from metonymies, but coordinated constructions are not the only scenarios where the literal and a metonymic sense are predicated together, or copredicated. *Copredication* is the phenomenon whereby the literal and metonymic appear simultaneously. For instance, Asher (2011) describes copredication in cases of conjunctions, where each argument has a different

semantic type. Conjunctions are but one of the structures that can lead to both senses being active at the same time.

(2.7) a) The Allies invaded *Sicily* in 1945.
b) We had a delicious leisurely *lunch*.
c) The case of *Russia* is similar.

Some verbs take arguments (cf. Section 2.5.2) that require both senses active at the same time and are known as *gating predicates*, following the claim by Rumshisky et al. (2007) that "there also seem to exist gating predicates whose selectional specification may specify a transition between two simple types". Thus, a verb like *invade* is a geophysically delimited military action, which requires both the LOCATION and the ORGANIZATION sense. Compare for instance with "Mongolia declared war against Japan", where only the ORGANIZATION is active for both *Mongolia* and *Japan*.

Also, there are contexts in which different elements affect the sense of the predicated noun towards being literal and metonymic at the same type without being coordinated. In b), the adjective *delicious* selects for the FOOD sense of lunch, while *leisurely* activates the sense of lunch as an EVENT, as only things that happen can be leisurely.

Example 2.7.b) also shows that metonymic senses can be propagated through anaphora, as the literal referent of the metonymy is maintained in the comprehension of the metonymic predication. For more on the ability of metonymies to preserve referent, cf. Nunberg (1995), Stallard (1993) or Asher (2011).

The last example has the word *Russia* placed in a context that does not indicate a strong preference for either sense. The copula has little pull towards either sense, and the lexical environment (*case, similar*) is also very vague. Without more context the sense of *Russia* cannot be resolved. It could be claimed that instead of having the two senses active at the same time like in a case of copredication, contexts of this kind have both senses *inactive* at the same time, achieving the same result, namely an underspecified reading for the potentially metonymic word.

Whenever a predication of a noun is potentially figurative and literal at the same time, we will refer to it as *underspecified*, regardless of the cause of such underspecification. In this way we group under the same term cases like copredication, gating predicates, vague contexts and the presence of multiple selectors, exemplified in 2.6 and 2.7.

The term underspecification has been used within the field of formal semantics to refer to e.g. the different possible analyses of quantifiers under a common representation (Egg et al., 1998). In this dissertation, the term will strictly be used for the predications of nouns that potentially manifest both literal and metonymic senses. The word *underspecified* can be misleading, however. By saying that the sense is underspecified, we do not mean that it cannot be resolved, but rather, that it can be both the literal and metonymic. Our usage of this term is borrowed from Cruse (Cruse, 1986, p. 153).

2.5 Generative Lexicon

The Generative Lexicon or GL Pustejovsky (1995) is a theoretical framework for lexical semantics developed to give account for the compositionality of word

meaning by improving over the sense-enumeration paradigm when building lexicons and thus offer a bridge between lexical semantics, predicate logic, and NLP.

The GL has been very influential both in theoretical lexical semantics and in natural language processing (cf. Section 3.2.1 for NPL applications that use the GL). The GL introduces a series of theoretical objects like *qualia structure*, *event structure*, and *dot type* to describe lexical entries, and how words described by these terms are predicated. The notion of dot type is central to this dissertation and it has its own section in 2.5.1; the other notions are briefly described in this section.

Pustejovsky offers a breakdown of the semantic traits of a word in four *structures* (from Pustejovsky and Ježek (2008)):

1. Lexical Typing Structure: giving an explicit type for a word positioned within a type system of the language
2. Argument Structure: specifying the number and nature of the arguments to a predicate
3. Event Structure: defining the event type of the expression and any subeventual structure it may have
4. Qualia Structure: a structural differentiation of the predicative force for the lexical item

a) Formal: the basic category which distinguishes the meaning of a word within a larger domain;
b) Constitutive: the relation between an object and its constitutive parts
c) Telic: the purpose or function of the object, if there is one;
d) Agentive: factors involved in the object's origin or 'coming into being'.

According to (Pustejovsky, 2006) here are two basic general types of nouns or *simple types*, with either two or four available qualia roles:

1. Natural Types: denoting nouns of natural kinds that have only Formal and Constitutive qualia roles available, like *tiger*, *river*, *rock*

2. Artifactual types: nouns for objects that have a purpose or have come into being, having thus Telic and Agentive roles, e.g. *knife*, *policeman*, *wine*

Qualia roles are not explicitly defined in the GL to be concepts or words and they are portrayed using a notation that resembles predicate notation in formal semantics (cf. Figures 2.3 and 2.4). Unless we have to interpret another author's lexical entry like in Section 2.6.2, we will refer to the values of qualia roles as being words. Understanding them as words also allows us to call them by their part of speech. The formal and constitutive quale are nominal, and the agentive and telic are verbal.

Lexical entries in the GL are normally portrayed as typed feature structures. In Figure 2.3 we can see the GL lexical entry for the artifactual noun *sandwich*.

$$\begin{bmatrix} \textbf{sandwich(x)} \\ \text{CONST} = \{\textbf{bread,...}\} \\ \text{FORMAL} = \textbf{physform(x)} \\ \text{TELIC} = \textbf{eat(P,w,x)} \\ \text{AGENTIVE} = \textbf{make\_activity(z,x)} \end{bmatrix}$$

Figure 2.3: Example GL lexical entry

A sandwich is an artifact and it is *made* by someone, thus having a saturated agentive qualia role. Also, it is made with a purpose and its telic role is *eat*. Since it is a physical object, its constitutive and formal roles are also saturated. Pustejovsky (2006) offers a more detailed account on how artifactuals are different from natural types, but for our purposes it is enough to mention that artifacts have at least one of the verbal qualia roles (agentive or telic) saturated. Simple (natural or artifactual) types can be combined to issue complex types—or *dot types*—such as the ones described in Section 2.5.1.

2.5.1 The dot type

In the previous sections we have defined regular polysemy (Section 2.2), metonymy (Section 2.3) and underspecification (Section 2.4). In Section 2.5 we also define simple type. These are necessary notions to introduce the notion of *dot type*.

In the Generative Lexicon (GL), Pustejovsky (1995) proposes the theoretical object of dot type (or dot object). A simple way to define a dot type is as a semantic type (or semantic class) made up two types: a simple type that provides a literal sense, and the most frequent metonymic sense provided by another simple type. The usual notation for dot types is a bullet (•) joining the grouped senses.

For instance, given that the most frequent metonymy for CONTAINER is CONTENT, we can group them together in a dot type, which is a complex semantic type where its members are originally of the CONTAINER type but can behave as members of the type CONTENT. In this manner, a word belonging to the CONTAINER•CONTENT dot type shown in (2.8) , like *glass*, can show its literal sense, its metonymic sense or an underspecified sense that can be result of copredication, vagueness, gating predicate, etc. Dot types are sometimes called complex types to emphasize that they are reifications of a grouping of more than one simple type.

(2.8) a) Grandma's *glasses* are lovely.

b) There is about a *glass* left in the bottle.

c) He left my *glass* on the table, and it was too sweet. .

Some works within the GL like Pustejovsky (2001, 2006) offer examples of dot types. We reproduce the list of dot types from Rumshisky et al. (2007) in Table 2.1.

Dot type	Example words
ACTION•PROPOSITION	promise. allegation, lie, charge
STATE•PROPOSITION	belief
ATTRIBUTE•VALUE	temperature, weight, height, tension, strength
EVENT•INFO	lecture, play, seminar, exam, quiz, test
EVENT•(INFO•SOUND)	concert, sonata, symphony, song
EVENT•PHYSOBJ	lunch, breakfast, dinner, tea
INFO•PHYSOBJ	article, book, cd, dvd, dictionary, diary, email, essay, letter, novel, paper
ORGANIZATION•(INFO•PHYSOBJ)	newspaper, magazine, journal
ORG•LOCATION•HUMANGROUP	university, city
EVENT•LOCATION•HUMANGROUP	class
APERTURE•PHYSOBJ	door, window
PROCESS•RESULT	construction, imitation, portrayal, reference, decoration
PRODUCER•PRODUCT	honda, ibm, bmw
TREE•FRUIT	apple, orange, coffee
TREE•WOOD	oak, elm pine
ANIMAL•FOOD	anchovy, catfish, chicken, eel, herring, lamb, octopus, rabbit, squid, trout
CONTAINER•CONTENTS	bottle, bucket, carton, crate, cup, flask, keg, pot, spoon

Table 2.1: Dot types from Rumshisky et al (2007)

From the sixteen dot types listed in the table, we have chosen a subset of five of these dot types to conduct the study of this dissertation. For a list and a rationale of the chosen dot types, cf. Section 2.6.

2.5.2 Simple and complex type predication

It is commonly postulated that argument-accepting words, called *functions* or *selectors*, have selectional preferences that establish the semantic type of their argument slots (cf. Měchura (2008) for a survey on selectional preferences). For instance, the verb *eat* requires an ANIMAL (including a person) as a subject and FOOD as an object. Using a noun of a mismatching semantic type into an argument role, thus forcing the mismatching word to be interpreted as a member of the desired type of the argument-accepting word is known as *coercion*.

In a sentence like "I will leave after the sandwich", *sandwich*, which belongs to the FOOD type is coerced into an EVENT by the word *after*, which places it in time. The word *sandwich* acquires duration as its meaning is extended to become "eating a sandwich". Pustejovsky (2006) breaks down the possible (mis)matching of types between functions and arguments in the following categories:

1. PURE SELECTION (TYPE MATCHING): the type that a function requires is directly satisfied by the argument

2. ACCOMMODATION: the type a function requires is inherited by the argument

3. TYPE COERCION: the type a function requires is imposed on the argument type. This is accomplished by either

   (a) EXPLOITATION: taking a part of the argument's type to satisfy the function;

   (b) INTRODUCTION: wrapping the argument with the type required by the function.

Introduction, the case for our *sandwich* example, is the most conspicuous case of coercion. Notice however how in this example, the interpretation of "after the sandwich" requires the activation of the telic role of the sandwich (*eat*). Even coercion relies on the activation of the verbal quale, e.g. "begin the book" exploits the telic *read*, "start the fire" introduces the agentive *ignite*.

Type exploitation is a characteristic phenomenon behind the predication of dot type nominals when only one of the possible senses is predicated and thus only a part of the arguments type is necessary to satisfy the function. Let us examine two simple examples for the dot type ARTIFACT•INFORMATION word *book*:

(2.9) a) I dropped the *book*.
b) I enjoyed the *book*.

These examples are complementary cases of exploitation. In the first example, *drop* only requires the book to be a physical object. In contrast, *enjoy* selects for the INFORMATION sense of the dot type. Both examples predicate a part of the full predicate force of *book*.

In terms of qualia structure, the verb *drop* activates the constitutive qualia role, which is the word that describes that books are physical objects made of pages and covers. The verb *enjoy* selects for the telic qualia role of *book*—that is, the function of books, which is to be read. This partial predication that happens with the exploitative type of coercion is what the GL uses as a case to postulate qualia structure as four separate roles and dot objects as complex types made up of more than one simple type.

The general picture is however more complicated. When predicating the INFORMATION sense of *book* with the verb *enjoy*, *book* is also placed in time as an event, like *sandwich* in our previous example, because what is meant to be enjoyed is "reading the book", which has a certain duration. There is thus a parallel saturation of the Event Structure described in Section 2.5 that happens along with the predication of the INFORMATION sense of *book*.

In addition to this, the coercion of non-event nouns into EVENT is very productive, and has spawned a subfield of research in its own right (cf. Section 3.3). It is for these reasons that we decide to abstract event coercion (both exploitation and introduction) away from our work by not dealing with event coercion of simple-type words (like *sandwich*) nor with the predication phenomena of dot-type nouns that are listed by (Rumshisky et al., 2007) as having EVENT as their first sense (cf. Section 2.6) .

The only exception to our abstention from dealing with eventive reading is that we incorporate the PROCESS•RESULT dot type into our study (cf. Section

2.6) because its PROCESS is not caused by coercion. Moreover, we consider we cannot disregard the theoretical relevance of this sense alternation, which is also isolated from event introduction and the other kinds of event exploitation like "lunch took forever" because the first sense is the eventive one.

In a related approach, Copestake and Briscoe (1995) take an approach to representing regular polysemy using a featured syntax like HPSG, extended to support qualia structure. They also argue for the reduction of a lexicon by saying that "film reel" and "fishing reel" and that the entry for *reel* is just a blank container entry and not two senses, because the sense is modulated upon predication. Some other kinds of polysemy, they argued, are best represented as lexical rules called *sense extensions*, like the grinding operation that turns animal into meat. Their stance is that the unwillingness of sense-extension-based polysemy to copredicate calls for different senses and is a lexical rule.

2.5.3 Compositionality, meaning and sense

Compositionality of meaning is a key factor in the GL. *Compositionality of meaning* implies that the overall meaning of an expression is a function of the meaning of its parts as they are organized by the syntax of the expression.

If meaning is compositional, it is reasonable to remove as much as possible from the sense listing and compose meaning when words are fitted into an expression. But if meaning is the output of a function over the parts of an expression, where does sense reside?

Research on logical metonymy—i.e. event coercion—describes sense as emerging from the relation between an argument-accepting word and the argument-saturating noun, which is customary from the coercion view of regular polysemy. From the coercion view it is acceptable to say that metonymic senses appear at the relation between elements, because coercing words are considered to select for a certain sense.

Complementary to the compositional view, a strictly distributional understanding of sense (Hanks, 2004; Kilgarriff, 1997) questions the idea that individual parts have individual atomic senses.

Without taking any strong ontological commitment to what actually bears the metonymic sense in the predication of a dot type, we will refer to the dot-type noun as having or showing a certain sense. This somehow trivially intuitive assumption also allows us to focus solely on a headword during the sense-annotation and the sense-disambiguation tasks in Chapters 4 and 7 respectively, instead of annotating word pairs or other more complicated structures.

Moreover, it might seem we have been using *meaning* and *sense* as apparent synonyms, which needs clarification. *Meaning* is used from now on to indicate the general property of semanticity, namely that messages bear some information, whereas we use *sense* to define the specific denotation that a word conveys in a given context. This division becomes useful when discussing the quantification of individual senses in 2.5.4.

Note that we are using a coarse sense inventory that corresponds to simple semantic types like LOCATION or ANIMAL. In the following sections we might refer to a dot type being made up of two semantic types, or exhibiting two possible alternating senses as it best fits the topic of the section.

2.5.4 Against sense enumeration

On his work on predication, Kintsch (2001) notes that

> Most words in most languages can be used in several different ways so that their meaning is subtly or not so subtly modified by their context. Dictionaries, therefore, distinguish multiple senses of a word.

But it is difficult to define what a word sense actually is, and how to find the boundaries between related senses.

There is a long list of possible definitions that try to describe what word senses are; from referencialistic conceptions of sense that depend on the notion of denotation, to cognitive approaches that depend on concepts (cf. Escandell-Vidal (2004) for a survey on theories of meaning). And of course, if a certain unit is not easily defined, it is also not easy to quantify. However, our representations of word meaning are more or less refined lists of senses (cf. (Battaner, 2008) for a review on the lexicographic approaches to representing polysemy).

Pustejovsky (1995) defines the senses in a lexicon as *contrastive* or *complementary*. Contrastive senses are those only related by homonymy:

(2.10) a) Sweden lost the *match* three to one.
b) She lit the candle with a *match*.

Complementary senses are those related by (metonymy-based) regular polysemy (cf. Section 2.3.3).

(2.11) a) The *bottle* fell and broke.
b) We shared the *bottle*.

What Pustejovsky calls a *sense enumeration lexicon* is a lexicon wherein contrastive senses are isolated from each other in different subentries (cf. the different subentries for *match* in Merriam-Webster we mention in 2.1) and complementary senses are kept together in a set:

1. if $s_1, ..., s_m$ are contrastive senses, the lexical entries representing these senses are stored as $w_{s1}, ..., w_{sn}$
2. if $s_1, ..., s_m$ are complementary senses, the lexical entries representing these senses are stored as $w_{\{s1,...,sn\}}$

Pustejovsky acknowledges the convenience of this kind of representation where semantics is isolated from syntax but he provides three arguments about the inadequacy of the sense-enumeration paradigm when describing lexical entries:

1. The creative use of words
2. The permeability of word senses
3. The expression of multiple syntactic forms

These three phenomena make traditional sense enumeration sufficient to deal with contrastive ambiguity—i.e. homonymy—but not with "the real nature of polysemy", according to Pustejovsky. The GL offers the notion of a dot type

(cf. Section 2.5.1) to group together the senses within a word without having to put them in a sorted list or an unordered set, as long as they are productive and a result of regular polysemy.

With regards to the **creative use of words**, words can incorporate a potentially unbound number of nuances to their senses in novel contexts. Pustejovsky does not propose that there are indeed infinite senses as such—a claim we also find in Cruse (Cruse, 1986, p. 50)—, but that there are regularities that can be represented in the lexicon in a way that makes lexical entries more succinct.

(2.12) a) a *good* book.
b) a *good* umbrella.
c) a *good* teacher.
d) a *good* sandwich.
e) a *good* pizza.

In the examples in 2.12 we can say that a good teacher is one that for instance teaches well, and a good book is pleasant or interesting to read. Goodness in the case of sandwich and pizza is defined in terms of tastiness, which is something all the FOOD words have in common. If the only consistency we can identify in the senses of *good* is at the semantic-type level, *good* has at least as many sense as there are semantic types. This dependence on semantic types is parallel to a selectional-restriction based understanding of senses.

An alternative to overspecifying lexical entries with very specific senses is keeping lexical entries purposely vague, so that the sense nuance is resolved through pragmatics or world knowledge. Vagueness is also a common practice when building dictionary entries for e.g. relational adjectives like *professional*, which Merriam-Webster describes in its first sense as "of, relating to, or characteristic of a profession". These are three different senses collapsed into one sense, an attempt to mirror the vagueness of statements like "a professional decision", where it is unclear whether the decision is taken by a professional—say, a specialist in something—or about one's profession or career.

Pustejovsky argues that there is a way to have properly specified lexical entries for words like *good*. Briefly, what the representation for *good* has to give account for, is that the word selects for the telic qualia role of the noun it modifies. In this way, *good* becomes "good to read" (*book*), "good to hold the rain" (*umbrella*), "good to eat" (*pizza*, *sandwich*), etc. A supporting argument for this claim is that "?a good iceberg" or "?a good penguin" cannot normally be resolved because these nouns are natural kinds without a telic role (however cf. Section 2.6.1 for further remarks on the introduction of the telic role for animal nouns).

This argument is very convincing for the case of adjectives like *good* or *fast*, and insofar the literal/metonymic alternation of a noun can be attributed to the selection of the telic qualia role, it also holds for nouns that have been specified to be dot types (cf. Section 2.5.1). For instance, in the sentence "I finished the book", *book* has the INFORMATION sense that is selected by the activation of the telic qualia role (*read*) by the verb *finish*.

But other areas of nominal sense variation push the limits of the GL as a linguistic program. The PART FOR WHOLE metonymy (also called synecdoche) presents a particular practical problem:

(2.13) a) After a *head* count, we saw we had lost Anna.
b) There are four *mouths* to feed in my home.
c) Hundreds of fast *feet* have gathered today to run the Marathon.

In these three examples there are body parts that stand for a whole person. Parts of a whole are called *meronyms*, and the whole is referred to as the *holonym* (Cruse, 1986, p. 162). For instance, *car* is the holonym for the meronyms *wheel*, *seat* and *motor*.

Meronyms are stored in the constitutive qualia role of a lexical entry in the GL, which poses two problems. First, there is no direct access from the part to the whole, but only from the whole to the part. When interpreting any of the examples in (2.13), we have to somehow retrieve that *person* is the whole for *mouth*, but this is not encoded in the qualia structure of *mouth*, and we would have to examine all the constitutive structures for all the nouns to find the whole that *mouth* is a part of, and then somehow resolve which of the candidate holonyms is the right one, since many things have mouths beside people.

Keeping an inverse list so parts can be retrieved for wholes—as they now can be retrieved from the constitutive qualia role—and wholes from parts would either require an additional structure in the GL, or an additional computational means of retrieval of this inverse relation.

Moreover, another computational limitation in the GL's take on synecdoche is that the amount of meronyms is ideally unbound, because the amount of parts a *thing*—an animal with its organs, tissues, bones, etc., or a time span with its successive subdivisions—can have is formally unbound, which makes the constitutive qualia role potentially very large. There should be incorporated a generative means to slim down the size of full constitutive qualia roles, maybe by more deeply integrating GL entries with the semantic ontology that the GL explicitly mentions but only defines at its highest levels (cf. Pustejovsky (2001, 2006)), or by establishing a mechanism to only include the meronyms that have linguistic relevance.

With the two previous remarks in mind we consider the GL method to deal with synecdoche challenging to implement in actual computational systems, because it presents information accessibility and cardinality issues.

The verbal qualia roles take the lion's share of the attention from theorists. The telic role is the one that has received most of the attention in the work (Verspoor, 1997; Bouillon et al., 2002; Yamada et al., 2007) that mentions qualia structure, because this work has often been devoted towards the study of event coercion (cf. Section 2.5.2) and the eventual reading of a metonymic word is often triggered by the activation of its telic role. The agentive role, being another (less) event-triggering relation, has also been studied, whereas it seems the formal and constitutive roles have been passingly mentioned or overlooked.

It is possible that the nominal roles, namely the constitutive and formal qualia roles have received less attention because they were considered less challenging or already-solved research questions. At any rate we believe that the constitutive role needs further refinement, in particular to how it contributes to offering an alternative from the sense-alternation paradigm. For further remarks on the amount of real-world knowledge to incorporate in the constitutive role, cf. Jensen and Vikner (2006).

The dot type is indeed designed as theoretical object to tackle productive, compositional, sense variation. Still, the metonymic senses that are listed are

sometimes too coarse or ill-defined in terms of ontological type (cf. Sections 2.6.5 and 2.6.3).

With regards to the inconveniences causes by rigid sense representations, Pustejovsky claims that the fixedness of representation in sense-enumeration lexicons clashes with the **permeability of word senses**, in that enumerated complementary senses overlap in a way that provides no natural level abstraction for the representation of these two alleged senses of e.g. *bake*:

(2.14) a) CHANGE OF STATE: John *baked* the potatoes.

b) CREATION: John *baked* a cake.

One is a verb of creation and the other is a verb of change of state. These are certainly complementary senses and not contrastive, and Pustejovsky rejects giving them independent sense status, because he claims that the second example is a special cause of the first one, and it is not possible to guarantee correct word sense selection on the basis of selectional restrictions alone. However, Pustejovsky does not mention that cakes are artifactual words that are made and can license a CREATION reading (i.e. by the agentive role), whereas potatoes are natural-type words and their sense for *bake* is strictly a change of state.

Moreover, if we want to study regular polysemy, the most pertinent critique to this understanding of polysemous senses being grouped in a set is the lack of description of the relation between them. It would be possible to give account for the sense subsumption in Example 2.14 by refining the mechanism of selectional preference instead of trying to refine the lexicon that Pustejovsky proposes, but this would not solve the problem of not explicitly representing the precedence relation—if any—between the ARTIFACT and INFORMATION senses of *book*.

Furthermore, the permeability of word senses that are otherwise described as being orthogonal is the major concern of this dissertations. It is indeed the possibility of metonymic senses coappearing with literal senses what triggers underspecified readings.

With regards to the common practice of assigning different senses to a word based on **difference in syntactic forms**, Pustejovsky considers it is equally arbitrary, as there is little difference in the core meaning of verbs like *debate*, regardless of its syntactic realization:

1. John *debated* Mary.
2. John *debated* with Mary.
3. John and Mary *debated*.

The case for this third point is built entirely around verbs, which is understandable because they show most variation in syntactic realization. This argument is complemented with the GL's classification on arguments, which we do not incorporate into a discussion, since our work is centered on the sense alternations of nominals. Nouns are indeed argument-accepting (cf. Section 2.6), but in our study do not require the larger theoretical apparatus that verbs would demand. Our focus on nouns makes us largely disregard this third argument.

In this section we have covered the three critiques that Pustejovsky offers on the sense-enumeration lexicon and we have in our turn criticized the alternatives offered by the GL. We acknowledge the computational comfort provided by

postulating qualia structure to simplify sense enumerations and give account for meaning generativity—e.g. the telic activation in the examples with *good*—, but we also find the constitutive qualia role to be unwieldy to implement. Regarding the possibility of overlapping senses, we take on the dot type as a theoretical object that includes more than one sense without having to necessarily choose only one of them when predicating, thus yielding an underspecified reading.

The GL has indeed been used a theoretical framework to develop lexical knowledge bases. In Section 3.1 we describe how the original formulation of the GL was expanded and modified to implement the aspects we have commented in this section, in particular the representation of qualia roles—especially the constitutive role— and the constituting senses of a dot type.

As a closing remark for this topic we also have to mention that the GL is implicitly geared towards metonymy-based regular polysemy, and it seems there is no place for metaphoric senses in the complementary-contrastive distinction mentioned at the beginning of this section. Certainly conventionalized metonymy is the firmest ground to start addressing the building of lexical entries for generative compositional meaning, but there are plenty of cases of productive figurative speech that are cases of metaphor-based regular polysemy (cf. Section 2.3) that a computational method build using the current incarnation of the GL would have trouble analyzing.

2.6 Study scope

We have determined the underspecified sense as our object of study. However, in order to attempt to identify it, we need to choose a series of dot types to conduct our study. In this section we detail the dot types that we will work with along the dissertation. First, we suggest a grouping of the dot types from Rumshisky et al. (2007) in different categories according to the semantic properties of the literal and metonymic senses that make up the dot type. These groupings are shown in Table 2.2.

Group A lists three dot types where both the first and the second sense are abstract. The members of group B have in common that they all share EVENT as a first sense. Group C has only one dot type, the class for words like *book*. Group D has three members, which are all ternary as opposed to the other binary dot types, and involve an organization or human group within their senses. E is the group for the words *door* or *window*. In G we find four dot types which have in common that their second metonymic sense is a concrete physical object which is consequence or result of the entity denoted by the first sense. PROCESS•RESULT is not included in G because we have no guarantee that the RESULT sense will be something concrete like a fruit or a motorbike and it is assigned to a singleton group. Lastly, CONTAINER•,CONTENTS also is placed alone in a group.

According to Sweetser (1990), original senses tend to be more specific and concrete. Nevertheless, we see that this tendency is not extensive to all metonymies. For words of the type TREE•WOOD, the WOOD sense is an extension of the original TREE sense. In the case of *class*, however, it is difficult to establish the fundamental, literal sense. Some of the dot types are difficult to analyze, notably the ones made up of more than two simple semantic types, such as the types for *newspaper*, *class* and *university* in group D. These ternary types are

Group	Dot type
A	ACTION•PROPOSITION
	STATE•PROPOSITION
	ATTRIBUTE•VALUE
B	EVENT•INFO
	EVENT•(INFO•SOUND)
	EVENT•PHYSOBJ
C	INFO•PHYSOBJ
D	ORGANIZATION•(INFO•PHYSOBJ)
	ORGANIZATION•LOCATION•HUMANGROUP
	EVENT•LOCATION•HUMANGROUP
E	APERTURE•PHYSOBJ
F	PROCESS•RESULT
G	PRODUCER•PRODUCT
	TREE•FRUIT
	TREE•WOOD
	ANIMAL•FOOD
H	CONTAINER•CONTENTS

Table 2.2: Groupings of dot types

dense metonymies in the Nunbergian terminology, in that there is no clear synchronic directionality of the metonymy.

Nunberg also concedes that the difficulties of analysis for these words are "in large measure the artifacts of our theoretical approaches—the need to say which uses of a word are 'basic' and which are 'derived'". We acknowledge this inconvenience and discard the dot types that are listed as ternary. Cf. Section 2.6.4 for our binary recast of a ternary dot type.

Ostler and Atkins (1992) mention that regular polysemy is sometimes blocked by "the pre-existing topography of the lexicon", a phenomenon they call preemption ("?eat the pig" vs. "eat the pork"). They also postulate one single sense alternation from container to amount/contents, and we make no distinction between the contents and a measure of unit (cf. Section 2.6.6 for the implications for Danish).

When listing dot types, Rumshisky et al. (2007) note that some of them are "pseudo dots" that behave as such due to very usual coercions, without pointing at which of the listed types are of such kind. In our work, we want to abstract away event coercion from the analysis (cf. Sections 2.5.2 and 3.2) because it is a different kind of sense alternation that the one strictly described by the sense alternations that give name dot types. Moreover, we decide not to work with the dot types in group B, that start with an EVENT reading, like the dot types for *lecture*, *concert* and *lunch*. Notice how none of the dot types in Table 2.1 has the EVENT sense at the right side, where metonymic senses are placed.

Both alternating senses in the dot types in group A are abstract, and that makes them difficult to apprehend, and good candidates for being discarded from a first study with human annotators in order to prioritize more immediate

sense alternations. For similar reasons, we decide not to work with group E, the APERTURE•PHYSOBJ dot type. This alternation, sometimes called figure-ground alternation (e.g. in (Pustejovsky and Anick, 1988)) has received a good deal of attention within the theory of metonymy but, even though both the first and second sense have concrete denotations, we consider this dot type to be overly difficult for human annotation to provide any conclusive results.

Having discarded A, B, D, and E we have four groups left, namely C, D, F, G, and H. Three of them are singleton groups, and we automatically incorporate their respective dot type to our study. From group G, we choose ANIMAL•FOOD over the other three because it will be easier to obtain corpus data; TREE•FRUIT and TREE•WOOD overlap in their literal sense, which complicates choosing one over the other during preprocessing and annotation. Moreover, the ANIMAL•FOOD alternation is a staple of the theory of regular polysemy. We also prefer ANIMAL•FOOD over the other three because it has received a lot of attention in the theory by Copestake (2013) or Copestake and Briscoe (1995), among others.

In the case of PRODUCER•PRODUCT, we also face a more complicated scenario that the dot type implies, because all of the listed producers are company names which also experience metonymies like ORGANIZATION FOR HEADQUARTERS, ORGANIZATION FOR STOCK MARKET INDEX and so on.

Now we have chosen ANIMAL•FOOD, CONTAINER•CONTENTS, INFO•PHYSOBJ and PROCESS•RESULT for our work. But there is no dot type to give account for the word *England* in the examples at the beginning of this chapter. The closest to the metonymy LOCATION FOR PEOPLE or LOCATION FOR ORGANIZATION is the ORGANIZATION•LOCATION•HUMANGROUP dot type in group D, which we had discarded as unwieldy. We thus incorporate a last binary dot type for our study to give account for the names of places. We choose LOCATION as the fundamental sense and pool together the ORGANIZATION and the HUMANGROUP metonymic senses in one ORGANIZATION sense which includes both.

The names of dot types also experience terminological variation, which indicates different conceptualizations of the same phenomena. For instance, INFO•PHYSOBJ (Rumshisky et al., 2007) is sometimes inverted as PHYSOBJ•INFO (Pustejovsky, 2006). We decide to call this dot type ARTIFACT•INFORMATION because ARTIFACT is a more exact semantic type than PHYSICAL OBJECT. In the case of the dot type ANIMAL•FOOD we also prefer ANIMAL•MEAT for the sake of exactitude. Table 2.3 provides the chosen five dot types with our preferred terminological variant, example words and the abbreviations we will use to refer to the datasets annotated for a certain dot type in Chapter 4.

Dot type	Abbreviation	Example words
ANIMAL•MEAT	ANIMEAT	*chicken, lamb*
ARTIFACT•INFORMATION	ARTINFO	*book, novel*
CONTAINER•CONTENT	CONTCONT	*glass, jar*
LOCATION•ORGANIZATION	LOCORG	*England, Asia*
PROCESS•RESULT	PROCRES	*construction, portrayal*

Table 2.3: Chosen dot types, abbreviations and examples

Ostler and Atkins (1992) and Peters and Peters (2000) claim that each regular-polysemy sense alternation has a single clear base form. This implies

that, in dot object terms, one of the senses is the original, fundamental or literal, and the other sense is derived through metonymy. This directionality of the metonymy might not be as clear for each dot type (cf. Section 2.5.1) but we have isolated five in which we are able to take a stance on what is the original sense.

Dot type	First	Second	Temporal	Spatial	Mass/Count
ANIMEAT	concrete	concrete	x		x
ARTINFO	concrete	abstract	(x)	(x)	(x)
CONTCONT	concrete	concrete		x	x
LOCORG	concrete	abstract		x	
PROCRES	abstract	concrete	x		x

Table 2.4: Summary of features for the chosen dot types

Table 2.4 shows a series traits for the chosen dot types. The "First" and "Second" columns show whether the first (literal) and second (metonymic) sense is concrete or abstract. The "Temporal" column indicates whether there is a temporal or aspectual component in the sense alternation of a dot type, thus making a dot type exploit the temporal contiguity for its metonymic sense. Likewise, the "Spatial" column determines if the metonymy in a dot type is caused by physical contiguity.

The last column "mass/count" indicates whether there is a mass/count distinction in any direction between the first and second sense of the dot type. For instance, the metonymic sense MEAT has allegedly the distribution of a mass noun, whereas the metonymic sense RESULT behaves as a count noun.

2.6.1 The Animal•Meat dot type

The ANIMAL•MEAT dot type, abbreviated as ANIMEAT exhibits a count/mass alternation. Individual animals are subject to a linguistic (and physical) grinding operation that turns them into a non-individuated mass (Copestake, 2013).

Called "Mass/Count alternation" in the classic GL Pustejovsky (1995), this sense alternation's status as a dot type is questioned by Pustejovsky and Ježek (2008), who call it a "pseudo-dot". The cause for their cautious phrasing might be precisely the aspect component of this alleged dot type: if the animal stops being ANIMAL to be FOOD, its underspecified sense might be unlikely to be copredicated, thus weakening the reasons to consider it a full-fledged dot type.

Still, it is a pervasive class-wise phenomenon, because any edible ANIMAL has a FOOD sense in English, unless the FOOD sense is already lexically saturated by the few but frequent Norman meat names like *beef, veal, pork, venison*, or *escargot*. Ostler and Atkins (1992) refer to the phenomenon whereby a metonymy is inhibited by an already-existing lexical item as *preemption*.

However, the main reason to contest the status of ANIMAL•MEAT as a full-fledged dot type is that the metonymic sense is obtained by an introduction coercion (cf. Section 2.5). Animals are natural types and their telic role is ascribed at best, but it is at any rate not as intrinsic as the telic role *hit* is to the artifactual *hammer*—animals put a great deal of effort in avoiding fulfilling their ascribed telic role, which is *being eaten*. This means that the telic role in

the lexical entry for *chicken* is not saturated with *eat* as expected, but rather, the *eat* role is introduced when referring to animals as food.

With regards to the aforementioned issue of copredicating the ANIMAL and MEAT senses, the examples offered in Nunberg (1995) suggest that the literal and metonymic senses in ANIMAL•MEAT can be more independent that what the theory in the GL expects from a dot type:

(2.15) a) They serve meat from corn-fed (Arkansas, happy, beheaded) *chickens*.

b) They serve corn-fed (Arkansas, happy?, beheaded?) *chicken*.

In the second example, the FOOD sense of *chicken* accepts the *corn-fed* adjective, which refers to the way the ANIMAL was raised and can have an effect of the properties of the meat, but rejects other adjectives that seem to only apply to the living chicken in the non-metonymic paraphrase.

Studying the ANIMAL•MEAT dot type allows us to focus on the count/mass distinction and on a metonymic sense alternation that is achieved through the less usual temporal contiguity. If we determine that ANIMAL•MEAT shows too few underspecified examples, we can claim it is one of the "pseudo dots".

2.6.2 The Artifact•Information dot type

The ARTIFACT•INFORMATION dot type (or ARTINFO is commonly provided as an example of dot type, and it is often showcased as the prototypical case of dot type (Pustejovsky, 1995, 2001; Ježek and Lenci, 2007).

Figure 2.5 shows how words like *book* are portrayed as examples of double inheritance, because they are at the same time an artifact made of pages and an amount of information (the Mental category in the ontology of the figure).

If the convention is that the literal, fundamental sense of a dot type is provided first, there seems to be an understated contradiction in the way the ARTIFACT•INFORMATION dot type is postulated. In Rumshisky et al. (2007) and in Pustejovsky and Ježek (2008), the authors list respectively the dot type with the INFORMATION sense first or second.

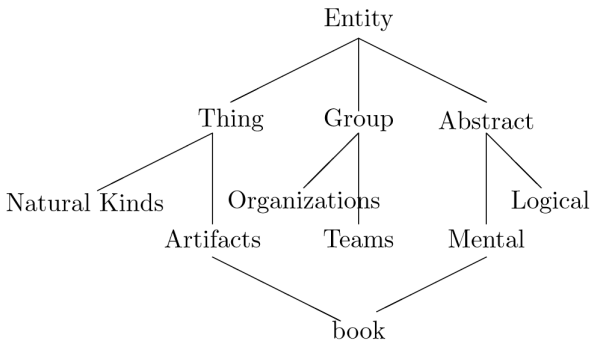

Figure 2.4: Multiple inheritance of ontological types for *book*

This ordering is certainly subject to interpretation: Is a book a mental entity that gets materialized by being written or printed, or rather, is it a physical

object that contains information? If the first case holds, the metonymy in ARTIFACT•INFORMATION is of the aspectual kind, but if the second interpretation is more adequate, then ARTIFACT•INFORMATION is a particular case of CONTAINER•CONTENT where the content is abstract and not physical. There might be competing understandings of what a book is that are either individual or circumstance dependent, making this metonymy one of the aforementioned dense metonymies. A writer might conceptualize a book using the first sequence (from thought to matter) and a reader might do the opposite.

Since there are more readers than writers, we choose to place the ARTIFACT sense first. It is also consistent with the rest of the theory that the activation of the telic quale (*read*) triggers the metonymic sense, which in our selected ordering is INFORMATION.

$$\begin{bmatrix} \textbf{book} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \texttt{y:information} \\ \text{ARG2} = \texttt{x:phys\_obj} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{FORM} = \texttt{hold(x,y)} \\ \text{TELIC} = \texttt{read(e,w,x.y)} \\ \text{AGENT} = \texttt{write(e',v,x.y)} \end{bmatrix} \end{bmatrix}$$

Figure 2.5: GL lexical representation for *book*

Figure 2.4 shows the lexical entry for *book* from Ježek and Melloni (2011). Its two verbal qualia roles, agentive and telic, require an argument of the type $x \cdot y$, where x is ARTIFACT sense and y is INFORMATION. In the GL, the verbs *read* and *write* are considered gating predicates because they select for the full dot type with both senses at the same type. The rationale behind this is that writing is the process of giving written transcription to an idea, and that reading is obtaining that idea from its physical support.

Besides the "Spatial" and "Temporal" values for this dot type being marked as uncertain in Table 2.4, the mass/count distinction is also unclear at first glance. Authors like Pinkal and Kohlhase (2000) provide examples where the ARTIFACT is a count noun and the INFORMATION is a mass noun, suggesting that this is a token/type distinction:

(2.16) a) Mary burned every *book* in the library.

b) Mary understood every *book* in the library.

In the second example the active sense is INFORMATION, which is a type (not token) reading. Asher and Pustejovsky (2005) do not dismiss the validity of the claim of a parallelism between the mass/count and the type/count distinctions, but note that such distinction does not explain how copredication works, and that type and token distinctions can be indeed predicated together:

(2.17) a) John hid every Beethoven 5th Concerto *score* in the library.

b) John mastered every Beethoven 5th Concerto *score* in the library.

It seems the much-studied ARTIFACT•INFORMATION dot type still offers plenty of challenges for empirical study to validate the formalisms that try to apprehend it.

2.6.3 The Container•Content dot type

In comparison to the other two previous dot types, CONTAINER•CONTENT comes across as less interesting or, at least, less challenging. This dot type can however not be disregarded because it is both the most immediate physical-contiguity metonymy of the chosen repertoire of dot types, and because this metonymic sense alternation is a source of prototypical examples of regular polysemy.

It is worth mentioning that CONTENT is not a sense in the way we have defined it in 2.5.3. Containers are made often with a certain content in mind, but the wine in a bottle, the bullets in a clip and the sardines in a can only have in common that they are physical entities. Ježek and Quochi (2010) use this semantic type more strictly and refer to it as LIQUID, but we are also using words like *box* or *can*, which refer to objects that can contain solids. Withal, we will refer to the metonymic reading of containers as the CONTENT sense for the sake of convenience in the following sections.

Additionally, the CONTENT metonymy can further be extended to the point that it becomes a measure of unit, which can later be conventionalized. The Imperial System uses *cup* as a measure of volume, and we make no distinction between the what-the-container-has-inside and measure-of-unit sense, because they are two nuanced manifestations of the same metonymy.

2.6.4 The Location•Organization dot type

As explained above, the interest of the LOCATION•ORGANIZATION dot type, which we propose as an alternative to the more complex types for words like *university*, is largely practical. The frequency of the metonymies of location names has impacted the field of Named Entity Recognition or NER (cf. Johannessen et al. (2005) for the different ways of addressing location-names metonymy for NER) and has spawned a line of research on automatic resolution of metonymies (cf. 3.4).

In considering the metonymic sense as abstract, we are focusing on the ORGANIZATION sense. However, for postulating this dot type we have merged together the ORGANIZATION and HUMAN GROUP senses into one single sense we have also called ORGANIZATION. It is worth noting that a human group would be a concrete noun, like a herd or a flock, while the ORGANIZATION sense is a third-order entity (cf. Section 2.3.2). Nonetheless, we expect the ORGANIZATION meaning to be more frequent than the HUMAN GROUP and take it as representative for the metonymies for LOCATION.

We choose this dot type because it represents another very usual metonymy, expected to be fairly easy to annotate—i.e. provide high-agreement datasets—and is relevant for NLP. This dot type is the only one whose datasets will be entirely made up of examples that have proper names as headword (cf. Chapter 4). Even though we assume the words *city*, *country* or *continent* have the same sense alternation, we have chosen only names of cities, countries and continents for our study to keep them in line with work in NER.

2.6.5 The Process•Result dot type

The sense alternation aggregated in the last chosen dot type, PROCESS•RESULT, has been a point of interest in lexical semantics since Grimshaw (1991) covered it in her studies on argument structure. Grimshaw claims a systematicity in the process/result alternation that is necessary for polysemy to be regular.

Even more so than in the case of CONTAINER•CONTENT (cf. Section 2.6.3), the metonymic sense RESULT is not a proper sense and it would not be possible to assign it to a single position in a semantic type ontology, something that is possible with other of the senses in the chosen dot types, like INFORMATION, FOOD, or ORGANIZATION. For instance, the resultative senses of the words *expression, construction* and *destruction* are even less related to each than the metonymic senses for *bottle, clip* and *can* compared in the section for CONTAINER•CONTENT. Resultative readings, however, have defined morphosyntactic traits, like being count nouns.

The work on distinguishing the PROCESS from the RESULT of nominalizations has been placed along the axes of determining whether a nominalization requires, accepts or refuses an argument, and what is the semantic difference of that argument being saturated or not. This is not an easy endeavor, and Grimshaw notes that:

> Nouns can and do take obligatory arguments. This property of nouns has been obscured by the fact that many nouns are ambiguous between an interpretation in which they take arguments obligatorily and other interpretations in which they do not.

Grimshaw's remark implies that there are arguments that are obligatory for a certain sense to be licensed but not for the other. In our discussion we will not differentiate arguments from being obligatory or optative (cf. Section 2.5), but rather on being present or not (cf. Section 6.1). By the same token we will not use Grimshaw's differentiation between complex and simple event nominals.

This dot type is a further oddity in that it the only dot type of this study whose first, literal sense is abstract and not concrete. In general we expert dot types with at least one abstract sense (and this dot type also has potentially abstract metonymic senses) to provide less reliable results when obtaining human-annotated data.

2.6.6 Choosing dot types for Danish and Spanish

Regular polysemy is allegedly a pervasive phenomenon in all languages (Apresjan, 1974), and there is no reason to believe that metonymy and metaphor are endemic of certain cultures, particularly from the experientialistic viewpoint of Lakoff and Johnson (1980), which considers metaphor and metonymy as fundamental strategies of conceptualization each human has available. Empirical work on NLP also treats metonymy as a cross-lingual phenomenon (Peters, 2003).

However, metonymy can be spotted in morphosyntactic alternations like count noun versus mass noun, for instance. These grammatical alternations are bound to be language-dependent, and, like Copestake (2013) mentions, some words show regular polysemy in one language because they experience

a metonymy, whereas in another language the same sense alternation is expressed by different words, and it is not possible to refer to their relation as polysemy. Apresjan also comments that the mechanisms for word formation are fairly similar to sense extension, only the former assigns new sense to a new item, while the later assigns new sense to an existing item.

For instance, the TREE•FRUIT dot type is for instance not to be found in Spanish. In Spanish there is a lexical rule that relates a fruit and the tree that bears it, both words sharing the stem but the tree being masculine and the fruit being feminine, like *manzan**a*** (femenine, *apple*) vs. *manzan**o*** (masculine, *apple tree*). Insofar gender is an intrinsic morphological feature of the Spanish noun (unlike adjectives, which obtain their gender by agreement), *manzana* and *manzano* have different lexical entries, and we cannot talk of polysemy, but of word formation.

We expand our study of the dot types in 2.6 to Danish and Spanish, in the somewhat traditional approach of comparing Germanic with Romance languages. The theoretical objects in the GL have been used in work in more languages than English, including French (Bouillon et al., 2002), Italian (Johnston and Busa, 1996), Romanian (Dinu, 2010), Spanish (Castaño, 2013), Korean (Im and Lee, 2013) or Japanese (Ono, 2013).

Collecting human-annotated data is a very cumbersome process, and we limit the scope of the study in Danish and Spanish to the same two dot types per language. We choose two of the dot types as candidates for the study across languages, namely CONTAINER•CONTENT and LOCATION•ORGANIZATION.

The dot type CONTAINER•CONTENT is the most prototypical metonymy of the chosen five dot types, and we consider it a good candidate for the comparative study because it is present in both Danish and Spanish. On the other hand, LOCATION•ORGANIZATION is the most frequent of the dot types described in Section 2.6 and its sense alternation is also present in the studied languages.

When prioritizing the dot types for our study in Danish and Spanish, we downweighted the relevance of ANIMAL•MEAT because its status as dot type is not completely reassured. We also discarded the ARTIFACT•INFORMATION for the cross-linguistic study. This dot type can be conceptualized in at least two different manners as we discuss in Section 2.6.2, and it is a candidate for a dense metonymy, which makes it potentially more difficult to interpret. For similar reasons we also discard PROCESS•RESULT.

The typological differences can have an impact in how dot type nominals are predicated and how their underspecified sense can be conveyed. Spanish does not require explicit pronominal subjects and presents a system of verbal clitics that can relate to objects, thus making the anaphoric factors of metonymy more likely.

On the other hand, Danish has the most complex nominal inflection of the three languages in the study. Danish nouns can be inflected for definite/indefinite, regular/plural and genitive/unmarked, which means each noun has potentially eight forms. See Example 2.18 for the possible inflectional forms of *hest* (*horse*).

(2.18) a) *hest*: indefinite singular unmarked (*horse*)
b) *hest**e***: indefinite plural unmarked (*horses*)
c) *hest**en***: definite singular unmarked (*the horse*)
d) *hest**ene***: definite plural unmarked (*the horses*)

e) *hest**s***: indefinite singular genitive (*horse's*)
f) *hest**es***: indefinite plural genitive (*horses'*)
g) *hest**ens***: definite singular genitive (*the horse's*)
h) *hest**enes***: definite plural genitive (*the horses'*)

This difference will only be manifest in CONTAINER•CONTENT, because the proper names in the LOCATION•ORGANIZATION data for Danish will only present genitive/unmarked case alternation but won't have definite or plural marker. Furthermore, Danish, like other Germanic languages, has a defined syntactic structure for the metonymic "measure of unit" reading of CONTAINER words, which is bound to emerge during the annotation. A more detailed study is provided in Section 9.3.

From all these considerations follows that our human annotation tasks and the following NLP experiments will be centered on the following nine datasets:

Language	Abbreviation	Example words
English	ENG:ANIMEAT	*chicken, lamb*
	ENG:ARTINFO	*book, novel*
	ENG:CONTCONT	*glass, jar*
	ENG:LOCORG	*Germany, Africa*
	ENG:PROCRES	*construction, portrayal*
Danish	DA:CONTCONT	*glas, kande*
	DA:LOCORG	*Tyskland, Afrika*
Spanish	SPA:CONTCONT	*vaso, jarra*
	SPA:LOCORG	*Alemania,África*

Table 2.5: Final nine dot types in three languages with study and abbreviated name for their corresponding dataset

2.7 Dissertation overview

In this chapter we have given an overview on the theoretical framework of this dissertation. We have defined the key concept of regular polysemy, how it relates metonymy, and how it allows postulating the dot type.

After providing an introduction to the different linguistic structures postulated in the Generative Lexicon, we have given a definition of the—also a key concept—dot type and how it is a grouping of a literal sense and its most frequent metonymic sense, enriched by the GL structures of event, argument and qualia structure.

Once we had a definition of dot type, we have analyzed a canonical list of dot types and chosen five them for a study in English, and two out of these five for a contrastive study that also includes Danish and Spanish.

We started the introduction with the question whether the dot type is an adequate useful means to represent regular polysemy at the token level for NLP tasks. In this chapter we have defined regular polysemy (cf. Section 2.2), and the notions of metonymy 2.3.1, but most importantly, we have defined sense underspecification in dot-type predications (cf. Section 2.4).

Underspecified usages of dot type words are those instances where neither the strictly literal or metonymic sense fully covers the conveyed sense of that

instance. With a definition of underspecification available, we specify the initial question in a more formal manner: Does sense underspecification justify incorporating a third sense into sense inventories when dealing with individual dot-type predications?

Proving the formal validity of the underspecified sense is fairly straightforward; take a copredicated example. Choosing only the literal or metonymic sense for an example is "England is rainy and organized" is impossible. However, this is a synthetic example, a common fixture in the literature of lexical semantics.

Instead of using synthetic examples, we want to assess the **empirical evidence for the inclusion of an underspecified sense** in our sense inventories. We address this question from three different angles:

1. We analyze the **human judgments** on sense underspecification in Chapter 4. In order to do this, we collect human annotations for corpus examples of the selected dot types. If human annotators cannot consistently annotate the underspecified sense, its applicability to NLP tasks is to be called into question.

2. We analyze the **distributional behavior** of dot type words using word sense induction in Chapter 5. Finding a distributional means to capture the underspecified sense would be sufficient to postulate its inclusion in our sense inventories.

3. We analyze the performance of a word-sense disambiguation system to predict the literal, metonymic and underspecified senses in Chapter 7. Achieving a supervised method that can **reproduce the human judgments** for the underspecified sense would be sufficient to postulate the inclusion of the underspecified sense in our sense inventories.

Note that our goal is not to change the way regular polysemy is represented in lexical knowledge bases, but to study the representation of the sense at the token level. Lexical knowledge bases that cover regular polysemy can express the literal and metonymic senses as a sense enumeration, or as a lemma-wise or class-wise relation between semantic types, but do not have to provide a representation for the underspecified senses, or for the senses that are *somewhat in between* the fully literal and fully metonymic readings.

Chapter 3

State of the art

In Chapter 2 we have provided an account of the concepts necessary to define our main research question, namely whether there is substantial empirical evidence to include the underspecified sense in sense inventories for dot-type predications. We have described regular polysemy, how it is often caused by metonymy, and how a set of metonymic alternations is grouped under dot types. Furthermore, we have delimited the dot types that our study will incorporate.

Most of the sources for Chapter 2 are works in linguistic theory of lexical semantics. In this chapter, we describe empirical approaches that make use of the concepts of such theory to identify metonymic senses from actual corpus data—and not from synthetic examples– using human or automatic means to discover them.

This chapter provides a revision of the previous work related to the general topic of study of this dissertation, namely the characterization of regular polysemy by humans and how to identify it automatically, particularly in the cases of underspecification. Empirical work on computational semantics has only so often dealt explicitly with regular polysemy (cf. Boleda et al. (2012a) for similar remarks). In this light, our take is to include in our literature survey the works that coincide with the topic of the dissertation by either being similar in object of study or method.

Each section covers one particular aspect of the empirical study of regular polysemy. Firstly we revise how lexical knowledge bases have addressed regular polysemy in their representations in Section 3.1. Section 3.2 describes annotation schemes that focus on capturing the difference between literal and figurative senses, particularly figurative senses caused by metonymy. In Section 3.3 we examine the existing attempts to model regular polysemy *at the type level* by using distributional semantics to identify the most likely sense of a potentially metonymic word pair or the different semantic types a polysemous word can belong to. Section 3.4 covers the attempts to model regular polysemy *at the token level* by disambiguating individual examples as literal or metonymic. In Section 3.5 we focus on a particular aspect of the works described in the previous sections to describe how they interpret disagreement between annotators as an indicator of polysemy. Lastly, in Section 3.6, we list the relevant design choices of the reviewed works that we incorporate in our human annotation and NLP tasks.

3.1 Lexical knowledge bases

The first step in dealing with figurative language is identifying and isolating the different senses words can manifest. Keeping a slightly artificial division between linguistic theory and state of the art, we have covered how this has been addressed from the theory in lexical semantics in Chapter 2. This section covers the representation of regular polysemy in lexical knowledge bases (LKB).

As Markert and Nissim (2002b) note, print dictionaries only include conventional metonymic senses, but metonymies are productive and potentially open-ended, and even the coverage of conventional metonymies is not systematic. Moreover, the example lists provided in the literature on regular polysemy are often synthetic, clear-cut examples that seem to oversimplify the distinction between literal and metonymic. Some LKBs explicit the relation of polysemous senses by links between words, while others do not include these relations.

A well-known LKB is Princeton Wordnet (Fellbaum, 1998). Wordnet takes no stance on whether the multiple senses of a word can be caused by regular polysemy because senses are listed as properties of words and not as properties of classes. The listing and representation of metonymic senses is not systematically addressed: not all (edible) animals have a FOOD sense, and the ORGANIZATION sense of locations can be found for the common nouns *city* and *country*, but not for the country and city (*Tanzania*, *Seoul*) names that appear in WordNet, which are hyponyms of *city* and *country* and should in principle inherit their sense alternation.

A family of LKBs built using the GL as theoretical framework is the SIMPLE lexicon (Lenci et al., 2000), designed to aid the development of wide-coverage semantic lexicons for 12 European languages. The SIMPLE lexicon provides qualia structures for its lexical entries and, more importantly, regular polysemous classes represented by a *complex type* which establishes a link between systematically related senses, thus covering the dot types described in Section 2.5.1. The Danish wordnet (Pedersen et al., 2006, 2009), based on the SIMPLE lexicon, holds regular-polysemy motivated relations between senses. Also, SIMPLE represents qualia roles as relations between senses, and not as words or abstract predicates.

Moreover, SIMPLE implements a PART-OF relationship from meronym to holonym—i.e. a branch is a part of a tree—to make up for the shortcoming of the formulation of the constitutive quale we describe in Section 2.5.4. In the original formulation, meronyms can be retrieved from the lexical entry of holonyms, but not vice versa; but by providing this additional relation the words can be retrieved from either side of the holonym-meronym relation.

Another field of work within LKBs is to automatically discover regular polysemous classes using unsupervised methods that follow the intuition that a sense alternation needs to be present in a plural number of words for that kind of polysemy to be regular, following claims like the ones by Dolan (1995) or Vossen et al. (1999) that conventional figurative senses are often represented in LKBs but not related to each other.

Peters and Peters (2000) exploit WordNet in hopes to find regular polysemy by looking at words whose hypernyms share sets of common words, thus finding class-based groups of polysemous sense alternations. They find pairwise combinations of classes that are numerous enough to postulate regular polysemy and that are mostly metonymic, like MUSIC–DANCE or CONTAINER–QUANTITY.

They argue for the productivity of the relations found, so that new WordNet entries for a certain semantic type would automatically have their regular polysemous senses available. They acknowledge however, that their approach does not obtain the directionality of a metonymy and cannot automatically determine which sense is literal and which sense is metonymic. Not knowing the original sense of a regular-polysemy sense alternation limits the automation of the productivity of regular polysemy. This shortcoming is pervasive all the automatic discovery techniques for regular polysemy described in this section.

Later, Peters (2003) expands on this work by trying to identify regular polysemy across languages (English, Dutch, Spanish) using the interlingual index provided by EuroWordNet (Vossen, 1998; Vossen et al., 1999). His results show that some patterns do indeed have a certain level of generality across languages, like for instance OCCUPATION–DISCIPLINE, e.g. *literature* is both something one can know about, and something one can work in.

Two related approaches to the automatic discovery of regular polysemy in LKBs are Buitelaar (1998) and Tomuro (1998, 2001b). Buitelaar developed CoreLex, a lexical resource defined as a layer of abstraction above WordNet that provides 39 basic types which work as ontological classes. Each class is linked to one or more WordNet *anchor nodes*, thus allowing for the multiple inheritance of properties for the anchor nodes' hyponyms. Tomuro's approach follows a similar motivation when she tries to find cut-off points in the WordNet ontology after which the senses are too fine-grained or too intertwined in their polysemous relation. The cut-off points in Tomuro's results are conceptually similar to Buitelaar's anchor nodes.

3.2 Sense annotation

This section describes sense-annotation tasks that have been geared towards capturing figurative senses, along with the annotation of concepts from the GL. We finish this section with a survey on the reliability of the results yielded by the reviewed tasks.

Collecting data for sense annotation allows a first evaluation of the adequacy of a sense representation. Thus, the empirical assessment of the theoretical object of dot-type underspecification does not start when measuring its learnability by automatic means, but when examining the results of the annotation task, even though the annotated data can later be used for machine learning.

Nissim and Markert (2005a) have extensively worked on human annotation for metonymic sense alternations. They summarize the principles and goals behind the annotation scheme that lead to further experiments (Markert and Nissim, 2006, 2007) and to the SemEval2007 metonymy resolution data Agirre et al. (2007). Their scheme is aimed towards the annotation of common nouns and proper names from the British National Corpus.

When annotating metonymies, the authors also work with the assumption of one sense being a basic, literal sense and the other senses being derived, metonymic senses. Each different semantic class—LOCATION, PERSON, ORGANIZATION, obtained in accordance with the MUC-7 Named Entity Recognition guidelines (Chinchor and Robinson, 1997)—can undergo a series of different metonymies and thus, each semantic class has its own sub-scheme listing the possible senses a word can manifest. They assume that the base semantic class

of the word is already known, that is, that metonymy resolution can be applied following Named Entity Recognition (NER) in a processing pipeline.

Their sense inventory is rather fine-grained. Instead of just providing one metonymic sense for geopolitical entities like we do in Section 2.6.4, Markert and Nissim attempt to annotate finer distinctions between types of metonymy. They provide for instance an OFFICIAL sense for a geopolitical entity ("*America* did try to ban alcohol") and CAPITAL-GOVERNMENT sense ("*Rome* decided"), which we find difficult to distinguish.

Moreover, they consider the metalinguistic usage of a name ("I think 'Elisabeth' is pretty") as a metonymy, and we consider it a homonymy, insofar the relation between a thing and its name is arbitrary, and does not depend on temporal or spatial contiguity. Also, they list object-for-representation, which we consider a metaphor in 2.3.1, as a metonymy.

More importantly, Markert and Nissim also incorporate what they call a "mixed reading", where "two metonymic patterns or a metonymic pattern and the literal reading are invoked at the same time". Our take on the mixed reading—what we refer to as *underspecified sense*—is different, because it is aimed at the simultaneous appearance of literal and metonymic senses.

In Section 2.3.3 we have shown the relation between metonymy and paraphrase. In accordance with this understanding, Markert and Nissim employ a replacement—i.e. paraphrase—test to assign the different metonymic senses. However, if two replacement tests are successful and one of them indicates the literal reading, they instruct the annotators to give preference to the literal reading. Also, they consider the possibility of no replacement being successful, which either issues a mixed reading, another kind of polysemy like metaphor, or an uncharted kind of metonymy that is not included in the sense inventory.

Their annotation scheme takes a strong position on the contexts that issue certain senses. The authors make a explicit listing on the syntactic contexts that can issue a mixed reading: two modifiers, a head and a modifier, the presence in a list, or being the subject of a verb complemented with a gerund or infinitive, as in "Sweden *claims to be* a great power". While this certainly improves the agreement if the data is annotated by experts, it makes it both biased towards a certain set of structures and difficult to implement using non-expert annotators. Instead of enforcing a certain sense to a certain structure, we expect these syntactic properties to emerge when correlating linguistic features to the sense of an example, and do not prescribe them from the start.

In a posterior account, Markert and Nissim (2009) provide an evaluation of the result of the annotation task that uses the annotation scheme we have reviewed, that is, the data for the SemEval2007 task on metonymy resolution. They note that metonymic senses tend to be skewed towards some very frequent metonymies, and they concede that there are many metonymic senses that are under-represented, making supervised learning difficult.

In spite of the differences in sense granularity, default sense backoff and the understanding of the mixed sense, this work is the main reference for the development of our annotation scheme. Besides taking in consideration their remarks on many metonymic sub-senses being under-represented, and thus collapsing them, we also assume Markert and Nissim's remark that their classification of metonymies by humans is hard to apply to real-world texts. The syntactic structure of sentences in corpora is in general more complex than in the very often made-up examples of the literature (simple clauses, maybe a coordination).

Pragglejaz (2007) proposes a metaphor identification procedure that has been used by Shutova and Teufel (2010) to annotate a corpus with verbal metaphors. Like Markert and Nissim (2009), they also annotate at the word level and not word relations.

The scheme by Shutova and Teufel also exploits a notion of basic meaning akin to our notion of literal sense (cf. Section 2.3.2), and they concede that basic meaning is not necessarily the most frequent sense of a lexical unit, a remark that Loenneker-Rodman and Narayanan (2010) also attribute to more theory-driven works like Nunberg (1995) and Hobbs (2001).

However, working with metaphors makes Shutova and Teufel's postulation of a fundamental sense easier than when working with metonymies, given that some words manifest dense metonymies (cf. Table 2.2), like words that belong to the LOCORG or the ARTINFO dot types.

3.2.1 Annotation from within the GL

Pustejovsky et al. (2009b) build on Markert and Nissim's scheme to introduce an annotation methodology which is based on the GL, and describe the Generative Lexicon Markup Language (GLML), to cover meaning composition upon predication. Since the scheme is developed to give account for phenomena like coercion or qualia activation, the authors recognize it is geared towards fairly complicated tasks.

Ježek and Quochi (2010) employ the GLML to annotate verb-noun coercions in Italian, and obtain a κ score of 0.87, which is very high score for a semantic task (and is on par with the results in Markert and Nissim (2007)). Two linguists (either Masters or PhD students) annotated each item, with one third senior annotator acting as judge and deciding in case of disagreements. The object of study of quite complex and they use qualified annotators, which also boosts their agreement over what it is expectable for a task of such complexity.

The GLML way of dealing with this kind of phenomena presents too much conceptual complexity to be crowsourced or annotated by volunteers with no formal training in linguistics. The authors also note that their scheme suffers from low agreement when the noun in the noun-verb coercion par is a metonymic noun or a dot type and present this problem as an open question. This makes us wary of using a coercion method to annotate the possible senses of dot types.

However, the GLML also contemplates the task of sense selection in dot objects. Pustejovsky et al. consider that, for a dot type that is modified by a verb or an adjective, either one of the two alternating senses is selected, or both are. These three options issue three possible senses for a dot type, namely the literal, metonymic and underspecified senses.

These coercion distinctions are very fine, and for instance, would only model the relation between the adjective and the noun in constructions like "delicious lunch" or "slow lunch". Thus, the only way to obtain an underspecified sense is when the noun is coerced by a gating predicate (cf. Section 2.3.3).

3.2.2 Reliability of annotation schemes

Human-annotated datasets are documented with agreement values to determine the reliability of the annotation procedure that generated them as a proxy for an

estimate of the reliability of the annotations. Cf. Section 4.6.1 for an overview on agreement measures and their interpretation.

Shutova and Teufel (2010) use three native English speakers as annotators. They report a κ of 0.64 in determining whether a verb is used in a literal or metaphoric fashion, and a lower value κ of 0.57 when trying to determine the source and target domain of the metaphors. They claim that the lower agreement in target-source domain identification is mainly due to the overlapping categories in their domain inventories—e.g. there is an IDEAS and a VIEWS domain, which are not disjoint.

Metonymy annotation is potentially easier that metaphor annotation, because there is only one domain involved. Markert and Nissim (2007) obtain very high agreement scores (A_0 of 0.94 and κ of 0.88), which is likely due to both annotators being not only experts, but also the authors of the annotation scheme, and because the authors previous fix the reading of certain syntactic structures. We would expect much lower agreement values if linguistically naive annotators were employed.

Boleda et al. (2012a) employed four experts to classify the 101 adjectives in their gold standard in the different semantic classes they wanted to automatically identify. They obtain κ scores between 0.54 and 0.64. In a previous work, Boleda et al. (2008) report a large-scale experiment to gather judgements on a semantic classification of Catalan adjectives, aimed at the study of the regular polysemy between adjectival semantic classes, and they allow multiple-class assignment for polysemy.

They enrol 322 participants in total and obtain rather low κ scores (0.20-0.34) but compare their annotation with an expert annotation and conclude that polysemous and event-related adjectives are more problematic than other types of adjectives. They also base their sense-annotation method in paraphrasing. Given their low agreement, they analyze the causes of disagreement and break it down in disagreement caused by the theoretical bias of their sense inventory and by the intrinsic difficulty of some sense distinctions (eventual adjectives seem the most difficult type).

Arnulphy et al. (2012) presents an annotation scheme for events which implicitly covers metonymic event readings in French and also obtain very high agreement (0.808 κ) when determining which words are events. Their scheme is followed by annotators who are also the authors of the very detailed guidelines. For other tasks, Merlo and Stevenson (2001) give κ values between 0.53 and 0.66 for expert classifications of verbs into unergative, unaccusative and object-drop. In a task closer to ours, Zarcone et al. (2012) obtain a κ score of 0.60 when determining which qualia role was selected for a certain metonymy of the style of "they enjoyed the cigarette" in German. Birke and Sarkar (2006) test their literal/figurative classification system on test data with a κ of 0.77

Variation in agreement coefficient is thus to be expected depending on the annotation setup, the kind of annotator, the sense inventory, the difficulty of the question that is asked to the annotator, and the difficulty of the examples in the corpus to annotate. Passonneau and Carpenter (2013) argue for a model of the labels of crowdsourced data to identify problematic labels as an alternative to providing an agreement coefficient.

3.2.3 Crowdsourcing data

Obtaining expert annotation is cumbersome and slow, and potentially expensive (in money or persuasion effort). It has become increasingly popular to employ Amazon Mechanical Turk (AMT) to crowdsource annotations for English for tasks like evaluation of translation quality, word sense disambiguation, event annotation, word similarity and text entailment (Snow et al., 2008; Callison-Burch, 2009; Rumshisky et al., 2009).

Using crowdsourced data comes often at the cost of reliability (Wang et al., 2013). Nevertheless, Snow et al. (2008) consider AMT able to provide reliable annotations for word senses provided that at least four annotators label each item, and they even find an error in the reference gold standard they were evaluating against by comparing it against the judgements of the annotators.

There has been an effort to obtain more consistent gold standard annotations by weighting the votes of annotators based on their agreement with a small expert-generated gold-standard or with the other annotators (Callison-Burch, 2009), or by automatically inferring a final label from all the labels assigned to an item using Expectation-Maximization (Raykar et al., 2010; Hovy et al., 2013).

3.3 Modeling of regular polysemy

Modeling regular polysemy has attracted attention because of the promising possibility of exploiting class-wise sense alternations in an unsupervised manner to build systems, and because it allows the empirical validation of lexical semantic theory.

Ježek and Lenci (2007), in their data-driven effort to characterize the GL-compliant semantic types of the possible arguments of the Italian verb *leggere* (*read*), emphasize the relevance of working with corpora to obtain models of word meaning. This section describes a series of NLP tasks that make use of distributional evidence to assess the behavior of metonymic words, i.e. that model regular polysemy at the type level, rather than the token level.

Lapata and Lascarides (2003) provide a statistical modeling of logical metonymy. Logical metonymy is a specialized variant of regular polysemy that is defined in terms of predicate-argument relations, that is, in terms of coercion, in particular event coercion (Godard and Jayez, 1993; Verspoor, 1997). The authors' concern is to characterize the syntax/semantics interface to incorporate a probabilistic element in the theory of composition of meaning. Even though we explicitly dismiss keeping our analysis to logical metonymy and event coercion in Section 2.5.2, this is a relevant article that sets the methodological groundwork for the empirical assessment of regular polysemy using distributional semantics.

In this work, Lapata and Lascarides provide the description of a series of probabilistic model that use distributional information extracted from a large corpus to interpret logical metonymies automatically without resource to pre-existing taxonomies or manually annotated data. They evaluate their results by comparing the system predictions with human judgements and obtain reliable correlation with human intuitions.

Their system works by comparing a metonymic construction's frequency ("enjoyed the book") with its expanded, literal paraphrase ("enjoyed reading

the book"). Obtaining the interpretation of a metonymic verb-noun pair is thus to retrieve which verb is implicitly meant in the coerced construction.

Their method consists of the modeling of the joint probability distribution of the metonymic verb, its object, and the sought-after interpretation. By doing this, they obtain verbs like *see*, *watch* and *make* as likely paraphrases for the metonymic verb-noun pair *enjoy film*. Their first model is centered around direct objects and does not take subjects in consideration.

They use human judgements to evaluate the quality of the induced paraphrases. They generate paraphrases with high, medium and low-probability interpretations ("Jean enjoyed the city" can mean he "enjoyed living in the city", "coming to the city", or "cutting the city"). Annotators were asked to rate paraphrases numerically according to their perceived plausibility. Their subjects could not have a background in linguistics. The Pearson correlation coefficient between the predicted likelihoods of the model and the human judgements is of 0.64.

They report inter-encoder agreement with leave-one-out resampling. The intuition is to eliminate one subject from the total m from the annotation data with $m-1$ annotators, calculate the correlation of their judgements, and repeat m times, each time removing a different annotator. They report an average human agreement in terms of correlation of 0.74.

Lapata and Lascarides' is a local model centered on interpreting coercion, and it does not take in consideration the overall context. Nevertheless, it is not a WSD method but a system that provides a ranking of the most likely interpretations of a verb-noun combination overall across all of its instances in the corpus. For instance, for *begin story* they obtain the following most-likely verbs for interpretation: *tell, write, read, retell, recount*.

Subsequently they incorporate subjects into their joint modeling to refine the obtainment of the interpretation verb: a cook begins a sandwich by *making* it, a client begins a sandwich by *biting into* it. In another experiment, and having incorporated both object and subject roles in their modeling, they aim to provide paraphrases for adjective-noun pairs, where the syntactic role of the noun (object or subject) in the paraphrase also has to be inferred. For instance, a "fast horse" is a horse that runs fast (subject), and "fast food" is a kind of food that is eaten or served quickly (object). This topic is a refinement of the coercion view and is further removed from our topic of research, but their agreement scores are much lower for adjectives (0.40 over the 0.74 obtained for verbs), something the authors attribute to adjectives having a wider range of interpretations than verbs.

Even though the their work is centered around coercion, it is similar to our understanding of how to capture metonymy using informants because they also make use of the paraphrase approach to interpret metonymies, and they claim that the interpretation of logical polysemies can typically be rendered with a paraphrase (cf. Sections 2.3.3 and 4.3).

The corpus-based modeling of coercion pairs has also been addressed by Murata et al. (2000), Shutova (2009); Shutova and Teufel (2010). In a recent work Utt et al. (2013) tackle the task from a different angle, namely by measuring the eventhood value of a certain verb, instead of determining the interpretation of a verb-noun pair.

A GL-compliant interpretation of Lapata and Lascarides' work is that obtaining the implicit verb in a metonymic coercion is qualia extraction of telic

and agentive roles. For other approaches to qualia role extraction, cf. Bouillon et al. (2002); Yamada et al. (2007) and a recent, general survey by Claveau and Sébillot (2013). For general acquisition of semantic types where regular polysemy is involved, cf. Bel et al. (2010, 2013) and Joanis et al. (2008).

Boleda et al. (2012a,b) describe work on modeling regular polysemy for Catalan adjectives at the type level by identifying sense alternations from corpus, generalizing above individual words, and predicting the sense alternation of new words unseen by the model during the fitting phase. In Boleda et al. (2012a) the authors experiment with two different models to determine the membership of a semantic class to a regular-polysemy alternation; one in terms of hard clustering where intrinsically polysemous classes are independent clusters (i.e. there is a cluster for sense A, a cluster for sense B and a cluster for senses A-B together), and a second one which models regular polysemy in terms of simultaneous membership to multiple classes.

Their second system fares best, and that leads them to assert that polysemous adjectives are not identified as a separate class. They obtain an accuracy of 69.1% over their 51% baseline, which they deem satisfactory given that the upper bound set by the agreement on their data is at 68%.

Boleda et al evaluate the practical and theoretical repercussions of the two different understandings of the relation of a polysemous adjective to its senses: the expectation that polysemous adjectives show a particular hybrid behavior and a "polysemous between A and B" can be postulated and identified as an atom—an expectation that has not been borne out—or the expectation that polysemous adjectives will show alternatively the behavior of one sense or the other. They claim that the treatment of regular polysemy in terms of independent classes is not adequate.

3.3.1 Modeling dot-object sense alternation

Rumshisky et al. (2007) offer the first—and to the best of our knowledge, only—regular polysemy modeling that explicitly addresses the study of dot types, in this case for English. This work also provides the canonical list of dot types we analyze and prune in Section 2.6. Rumshisky et al. obtain, in an unsupervised manner, the words (verbs and adjectives called *selectors*) that determine whether a dot type manifests one of the two alternating senses, or both.

Much like Lapata and Lascarides (2003), this is also centered around verb-noun and adjective-noun coercion. According to their results, the PROCESS•RESULT word *building* is a physical entity (RESULT) when it appears as the object of verbs like *demolish, enter* or *leave*, and when it is the subject of *stand, house* or *collapse*. The PROCESS sense is selected when *building* appear as the subject of *commence* or the object of *allow* or *undertake*. Most importantly, this work also finds gating predicates, words that select the whole dot type and thus issue the underspecified meaning, as we have defined in Section 2.3.3, by selecting simultaneously both alternating senses (PROCESS and RESULT in this example).

The authors obtain the selectors for each lemma in their list using a clustering method based on a particular kind of distributional similarity, where two words are similar if a certain argument (object or subject, for verbs) is often saturated by the same word. For instance, *cook* and *prepare* are similar because they often have *lunch* as an object. Once words are characterized in terms of their relation to nouns, they are clustered using agglomerative hierarchical clustering.

Interpretation of the clusters determines whether the words that appear together are selectors for the literal or metonymic sense. If they select for both, they are selectors for the underspecified sense, which Rumshisky et al. note by providing the full name of the dot type in question. The authors estimate that their clustering method yields more consistent clusters for verbs than for adjectives, which is in line with the higher reliability of the study on verbal coercion over adjective coercion in Lapata and Lascarides (2003).

This work is relevant for our purposes because it models the relation between selectors and the possible senses of the analyzed dot types, and typifies each selector as a feature for each of the three possible senses of a dot type in our sense inventory.

After a type-based modeling task, Boleda et al. (2012a) concludes that there is a need to move to word-in-context models, that is, token-based word sense disambiguation. However, using the association of argument slots of verbs to a certain sense as features à la Rumshisky et al. is a kind of coercion-based, selectional restriction understanding of metonymy that we are abstaining from.

In previous work, Boleda et al. (2007) treat the issue of polysemy as a multi-label classification problem. A word that presents regular polysemy will be tagged as being a member of more than one semantic class. This approach is used for type-wise modeling but can be extended to the token-level case.

3.4 Figurative sense resolution

The automatic, token-based identification of word senses in contexts is called *word-sense disambiguation* (WSD). It is a large discipline that encompasses a multitude of approaches and has spawned many shared tasks, from the early SenseEvals (Kilgariff and Palmer, 1998; Preiss and Yarowsky, 2001) to the last SemEval (Navigli and Vannella, 2013a; Navigli et al., 2013).

WSD has also contributed to the theory of lexical meaning and its relation with pragmatics, yielding the one-sense-per-discourse approach to word senses (Gale et al., 1992)—more apt for dealing with homonymy— and the one-sense-per-collocation approach (Yarowsky, 1993)—more apt for dealing with actual polysemy.

In this section we cover works in WSD that explicitly try to predict figurative senses, or the sense alternation of metonymic words. The application of WSD to metonymy has been called *metonymy resolution* Markert and Nissim (2002a).

We extend this term to *figurative sense resolution* for the cases of WSD that focus on the recognition of figurative senses regardless of their nature, be it metonymic or metaphoric. In this section we also refer to work that has been useful in building the feature model described in Section 6.

Figurative sense resolution is not confined to statistical machine learning, and from the early days of rule-based, symbolic applications to NLP there have been approaches that try to disambiguate word usages as literal or figurative (Hobbs and Martin, 1987; Onyshkevych, 1998; Stallard, 1993; Bouaud et al., 1996).

Authors like Stallard (1993) already argue for the necessity of resolution of metonymy for nominals, particularly as a preprocessing to make the entity replaced by the metonymy explicitly available for e.g. machine translation. Most of the works of the period like Fass (1988, 1991), for metaphor and metonymy,

and Markert and Hahn (2002) or Harabagiu (1998), centered strictly around metonymy, attempt to infer whether an expression is figurative by analyzing the path in a in a hierarchy of semantic types generated to interpret the expression.

In Kintsch (2001, 2000) we find an early attempt to use a distributional semantic model (Latent Semantic Analysis) to interpret the semantic properties of a noun user in "A is B"-style copulative metaphors like " my lawyer is a shark". Kintsch follows the two-domain notion to model the properties of noun B (*shark*) that get transferred unto noun A (*lawyer*). Although compelling, Kintsch's is system is constrained to synthetic examples of the form "A is B", which makes this algorithm difficult to apply to proper corpus examples, even more so than the coercion models described in Section 3.3, which are also based on modeling the meaning of word pairs. Moreover, this approach is tailored for metaphor and would not be immediately transferable to processing metonymy, where there is no transfer of properties across domains (cf. Section 2.3.1). Still, this is a seminal work on using distributional semantics for figurative language, which we also explore in Chapter 5.

Upon the establishment of statistical machine learning as reigning working paradigm in NLP, metonymy was identified as a source of noise for tasks like Named Entity Recognition (Leveling and Hartrumpf, 2008; Johannessen et al., 2005; Poibeau, 2006), which called for its automatic recognition.

Nissim and Markert (2003) understand metonymy as a classification task and use syntactic features to identify it. In order to generalize over the lemmas of the syntactic heads, they use an automatically generated thesaurus to cluster the syntactic head features. In Nissim and Markert (2005b) they mention the relevance of other features like the presence of determiner and the amount of syntactic dependencies the headword to disambiguate is involved in, plus they discourage the use of n-grams for metonymy resolution.

Markert and Nissim (2009) offer a summary of the results of the SemEval2007 shared task on metonymy resolution. They compare the results of the submitted systems against three baselines, namely MFS, SUBJ and GRAM: the usual most-frequent sense (MFS) baseline used in WSD, the SUBJ baseline where all the instances of potentially metonymic words that are subjects are labelled as metonymic, and the supervised GRAM baseline where each syntactic role receives its most likely sense as determined from the training data. The last two baselines are harder than MFS for their data, even though in highly skewed data MFS is a very hard baseline to beat. We analyze these three baselines in Appendix D.

Since they use a very fine-grained sense inventory, Markert and Nissim's evaluation was conducted using three levels of representation granularity: COARSE (literal vs non literal), MEDIUM (literal, metonymic or mixed) or FINE, with all the subtypes of metonymy included. The authors also identify mixed senses as potentially problematic for WSD.

In their report, Markert and Nissim analyze which kind of features each submitted system uses. All three top-scoring systems used head-modifier relations, although performance is not homogeneous across senses, arguably because syntactic roles also are skewed. Some systems used collocations and co-occurrences but obtained low results, and in general it seems that shallow features and n-grams are not desirable for this task. Even though the best systems used syntactic roles and head relations, none made explicit use of selectional restrictions, which means no system was implemented following a strict coercion view,

and we also decide to abstain from it. For full system descriptions for the SemEval2007 metonymy resolution shared task submissions, cf. Agirre et al. (2007)

The GLML, described in Section 3.2.1 has also made possible a shared task at SemEval2010 (Pustejovsky et al., 2009a), portraying verb-noun coercion in English and Italian. There was one participant to the task (Roberts and Harabagiu, 2010), which only submitted a run for English, obtaining however very good results (e.g. a precision of 0.95 for determining whether a verb is exerting a coercion on an argument, over a baseline of 0.69). This system used induced hypernyms from WordNet and parse-derived features.

Birke and Sarkar (2006) modify a word sense disambiguation algorithm to classify usages of verbs into literal and figurative, using just sentential context instead of selectional restriction violations or paths in semantic hierarchies. Other systems that detect figurative language at the example level are Mason (2004); Murata et al. (2000); Sporleder and Li (2009); Turney et al. (2011), or Nastase et al. (2012), who use both local (words in the sentence) and global context (corpus behavior over a series of semantic classes) as features for metonymy resolution with SemEval2007 data.

Markert and Nissim (2009) describe metonymy resolution as class-based WSD. Regular polysemy has been a motivation for applying class-based WSD to account for the literal and metonymic senses of whole semantic classes. This allows reducing the sense inventory and improving the performance (Vossen et al., 1999). Similar remarks can be found in other class-based WSD systems like Izquierdo et al. (2009); Ciaramita and Altun (2006); Curran (2005); Clark and Weir (2001); Lapata and Brew (2004) or Li and Brew (2010).

3.5 Measuring uncertainty

Some research in NLP treats uncertainty in annotation as a source of information, and tries to represent its predicted variables as continuous values instead of discrete categories to palliate the bias imposed by the discrete representation. In this section we provide a review on works that either use uncertainty of annotation as information, or that try to predict semantic characteristics as continuous instead of discrete values.

Sun et al. (2002, 2003) attempt to implement fuzzy logic systems to avoid providing hard set assignments to their classifiers, aiming to represent a functionalistic language perspective where linguistic categories are seen as gradients, and they argue for the need to develop soft computing strategies that do not immediately commit to hard classes, especially when the trained systems feed their output to another NLP system, because information is lost when discretizing into mutually exclusive classes.

Current data-driven methods allow the materialization of theoretical statements that claim that two opposing categories in linguistics are in fact poles of a continuum, like abstract vs. concrete or homonymy vs. polysemy. Utt and Padó (2011) model the increasing systematicity in sense variation from homonymy to (regular) polysemy to postulate a continuum between the two: homonymy is idiosyncratic of a word whereas polysemy is characteristic of a certain class.

Utt and Padó interpret high-frequency ambiguities between semantic classes as systematic and low-frequency as idiosyncratic, which places homonymy and

polysemy in a gradient. They also argue that their model is desirable because phenomena have no separation boundary and words can have both homonymic and polysemous senses, e.g. the word *match* in Section 2.1. Utt et al. (2013) explicitly address the modeling of metonymic behavior as a graded range because they find unlikely that there are two exclusive classes of verbs, namely metonymic and literal, and model the verbs degree of metonymicity.

A task related to predicting the literality of an expression as a continuous value is has been implemented by Bell and Schäfer (2013). Their take is focused on the compositionality of a compound, following the notion that compositional compounds are more literal: the meaning of *application form* is composed from the meaning of its parts, whereas *ivory tower* is neither made of ivory not really a tower but a way of referring to the academic world. They obtain human ratings of the literarity of a compound and train regression models to predict this value, taking into account factors like frequency of the parts, or the kind of semantic relation between head and modifier in the compound. Cf. Johannsen et al. (2011) for a similar task that predicts the compositionality of a word pair.

Another work that uses a continuous representation of literality is Turney et al. (2011), which conceptually places examples in a continuum from the absolutely literal to the fully figurative by saying that they degree of abstractness in a word's context is correlated with the likelihood that the word is used figuratively. However, the continuous representation is only a factor in their feature modeling for classification, because their final system employs binary classification to divide examples as either literal or figurative.

3.5.1 Exploiting disagreement as information

Disagreement between annotators is commonly seen as some kind of featureless noise. However, there are authors that take the stance that, when an item is annotated with low-agreement and the labels it gets assigned are disparate, there is a case for seeing this particular example as difficult to annotate. One of the causes for this difficulty can be regular polysemy.

Tomuro (2001a) provides an interpretation on disagreement between sense-annotated corpora. Her hypothesis is that, when humans disagree on senses of a word, there is an underlying relation between the senses, that is, most inter-annotator disagreement is explained by systematic polysemy.

Tomuro compares the sense annotations of two corpora, SemCor and DSO, and reports an average K between the sense annotations of the two corpora in their matching sentences of 0.264. She concedes however that a good proportion of the difference is a result of SemCor being annotated by lexicographers and DS by novices, but claims that these differences provide insight on sense distinctions that are easily misjudged. She is also careful to note that the inverse of her hypothesis does not work, and that systematic polysemy does not cause cause disagreement per se. Ježek and Quochi (2010) also remark that the nouns that cause most disagreement in the coercion annotation task are precisely dot objects.

Recasens et al. (2012), dealing with co-reference annotation, work with an intuition of near-identity in coreference similar to our understanding of senses being underspecified. They build an annotation task where the five annotators had to determine whether an expression was co-referent or not. They also allow an option for the annotators to inform whether they were unsure about

a particular item. They obtain perfect agreement for only half of their data. From their annotations data, they generate a final, aggregate measure s, bound in the interval (0–1]. High-agreement data provides values of s closer to 1. Their approach is relevant to our interest because they do not only use the agreement to quantify the reliability of the annotation, but to merge the discrete annotations of five annotators into a single, continuous label. They mention on their future work section the possibility of predicting such disagreement-based value, a task we take up in Chapter 9.

3.6 Summary

In this chapter we have reviewed the previous work on LBKs abd human annotation for figurative senses, and their modeling in type- or token-based methods. In this section we summarize the methodological contributions of the state of the art to the tasks that make up the experiments of this dissertation.

3.6.1 Designing an annotation task

In Chapter 4 we provide the description and the results of the sense-annotation task for dot types that attemps to identify the underspecified sense, as well as the literal and metonymic senses. In this section we summarize the design choices we impose on our sense-annotation task after the literature survey in Section 3.2.

The annotation scheme is largely based on the annotation scheme by Nissim and Markert (2005a). There are, however, several differences in our approach. Instead of enforcing a certain sense to a certain structure, we expect these syntactic properties to emerge when correlating linguistic features to the sense of an example, and do not prescribe them from the start.

Markert and Nissim's scheme is more detailed but also more biased by the author's expectations, and using a less detailed scheme that does not incorporate any syntactic information we potentially sacrifice agreement in favor of descriptivism, as opposed to a prescriptivism of sorts, where they a priori enforce readings on syntactic structures.

The sense inventory in Nissim and Markert (2005a) is very fine-grained, and the authors found that many low-frequency patterns performed poorly at WSD. We cannot do away with the underspecified sense, which is the focus of our study, but reduce our sense inventory to the affordable minimum of three categories to annotate: literal, metonymic and underspecified.

We annotate with three possible senses. Thus, the granularity of our sense inventory is similar to their MEDIUM representation, keeping in mind that Markert and Nissim's mixed reading is different to our encoding of the underspecified sense.

Most importantly, our take on the mixed reading—what we refer to as *underspecified sense*—is different. In Markert and Nissim's guidelines, a doubly metonymic sense is also mixed. Our sense inventory only lists one possible metonymic sense for each dot type.

With regards to the GL-based sense-annotation guidelines in Section 3.2.1, we abstain from using a coercion approach to sense selection. A method centered in pairwise word relations is not the desired for an enterprise like ours, were

we will be dealing with corpus examples of arbitrary complexity, and we have established we will model the predicated sense of a dot type at the word level and not at the relation level (cf. Section 2.5.3)

Coercion phenomena are a subset of the possible contexts that provoke metonymic readings, and we have decided to take a general approach that does not explicitly focus on coercion. In Section 5.3 we cover, for instance, the relevance of articles and plurals for determining the sense of dot types. Not focusing on coercion also entails that we will model the predicated sense of a dot type at the word level and not at the relation level.

Most of the work on figurative sense annotation in Section 3.2 used experts as annotators. In this dissertation we use Amazon Mechanical Turk to annotate the English datasets, and volunteers for the Danish and Spanish data.

Markert and Nissim backoff to the literal sense in case of disagreement. When assigning a final sense tag from all the possible annotations for an example, we explicitly avoid automatically backing off to the literal sense in our annotation scheme. This decision is justified and illustrated in Section 4.7.1.

Jurgens (2013) argues for the multiple tagging per item to boost the agreement of datasets. Nevertheless, we abstain from multi-tagging in our sense annotation tasks, even though we interpret the underspecified sense tag as a double tag in the ensemble classification step in Chapter 7.

We have also listed methods to attempt to infer the most likely tags in an unsupervised manner. In 4.7.2 we detail how we use MACE (Hovy et al., 2013) to assign sense labels in an unsupervised fashion, and how we compare the output of MACE with a theory-compliant majority voting sense assignment method.

3.6.2 Extracting Features

The works in type-based (Section 3.3) and token-based (Section 3.4) modeling of regular polysemy make use of different linguist features as explanatory variables for their machine learning experiments. In this section we provide a rationale for the features we incorporate in Chapter 6, and for the features we do not implement in our system.

Using selectional preferences (Rumshisky et al., 2007) as features corresponds to modeling metonymy as coercion, which we abstain from doing. We do, however, incorporate syntactic heads and the dependents of the nouns to disambiguate, for which (Markert and Nissim, 2009) report good performance.

Nissim and Markert (2005b) mention the relevance of other grammatical features like the presence of determiner and the amount of syntactic dependencies the headword is involved in. We also incorporate these features. Moreover, they discourage the use of n-grams for metonymy resolution, as these features damage the precision of the sense resolution. We do not incorporate n-grams in our feature space.

In order to generalize over the lemmas of the syntactic heads, Markert and Nissim use an automatically generated thesaurus to cluster the syntactic head features. We use a similar approach using Brown clustering in Section 6.1.

Coercion-based work also uses features we can adapt for our system. Even though the task in Roberts and Harabagiu (2010) deals with coercion, the features employed by the system are also applicable to the study of dot type senses: hypernyms from WordNet, besides parse-derived features. We also incorporate wordnet-based features to our feature space.

Nastase et al. (2012), who incorporate and global information features for metonymy resolution. We do so by including topic models and Brown-cluster features in our feature scheme, to give account for the distributional information that can be inferred from corpora in an unsupervised manner. Using unsupervised machine learning to generate features renders our method unsupervised.

3.7 Word-sense disambiguation

In Section 3.6.2 we have covered the features from the state of the art that we incorporate in our feature space. In this section we cover the design choices for the word-sense disambiguation system to capture the underspecified sense that do not deal with the feature representation of the examples, but with other methodological choices.

Using classification to predict the senses of the annotated data using linguistic features is the standard outline of most WSD systems. In this manner, the WSD system described in Chapter 7 does not deviate from the usual method for (semi-)supervised WSD. The annotation task is at the sentence level, which necessarily makes our WSD system a sense-per-collocation (Yarowsky, 1993) system. However, since literal and metonymic senses can appear together, we consider this approach more apt than the sense-per-discourse approach, which is more useful for e.g. the homonymy between *match (flammable stick)* and *match (sports competition)* than for metonymic alternations.

However, we understand the underspecified sense as a group of behaviors that include both literal and the metonymic sense, which means we vulnerate the assumption of mutual exclusion of senses which is customary in WSD (cf. Section 2.5.4).

Besides our treatment of the underspecified sense, the main particularity of our WSD system is that it is class-based. This means we expect the same sense alternations for all members for a semantic class—i.e. dot type— and train and test the data for all words within a dot type together, disregarding the different lemmas the headwords might have.

Boleda et al. (2007) claim that the treatment of regular polysemy in terms of independent classes is not adequate, and address type-based modeling of polysemy as a multi-label classification problem, which we also implement in our token-based disambiguation task using a classifier ensemble (cf. Section 7.4).

Note that, since the definition of underspecified sense in this dissertation is different to the mixed reading of Markert and Nissim, we do not use their data to train or test. In spite of this difference, our approach has a granularity similar to their MEDIUM representation.

3.8 Representing literality in a continuum

In Section 3.5 we have covered works that represent a semantic phenomenon as a continuum, instead of discrete categories. In Chapter 8 we propose a continuous representation for literality that is based on the range of metonymicity of a verb described in Utt et al. (2013), and on the degree of literality of a compound in Bell and Schäfer (2013) or a word relation in Johannsen et al. (2011).

We adapt their representation of figurativeness (or its complement, literality) as a continuous dependent variable in Chapter 8, using an approach that does not use a final sense tag, but a merge of the discrete annotations for each example into a single continuous value, as in Recasens et al. (2012).

3.9 Predicting disagreement

In Chapter 9 we propose a regression method to predict systematic disagreement, namely the variation of agreement in sense-annotated examples that depends on the linguistic features of the headword and its contexts.

This method is a continuation of the further work section of Recasens et al. (2012), which suggests attempting to predict disagreement-based values.

Chapter 4

Human judgments on regular polysemy

In this chapter we describe a sense-annotation task to obtain human judgments on the literal, metonymic and underspecified senses of dot-type words in English, Danish and Spanish. This chapter is an expansion on the work presented in Martínez Alonso et al. (2013).

The goal of this annotation procedure is, first and foremost, to assess the human judgments on regular polysemy, namely how humans perceive and annotate the differences between the literal, metonymic and—most importanly—underspecified examples.

This annotation task produces data that can be used as gold standard for the experiments in the following chapters. We need gold standard data to evaluate any system we want to develop to resolve regular polysemy. A gold standard is even more necessary for supervised learning, because we also need it for training. The lack of available sense-annotated gold standards with underspecification is a limitation for NLP applications that rely on dot types (Rumshisky et al., 2007; Pustejovsky et al., 2009a). In this chapter we cover the annotation procedure we have followed to generate such data [1].

This annotation scheme is designed with the intention of capturing literal, metonymic and underspecified senses, and we use an inventory of three possible sense tags. We call the first sense in the pair of senses 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, and ORGANIZATION is its metonymic sense. When a sense includes aspects of both literal and metonymic, we call it *underspecified*.

We annotate data in three languages—Danish, English, and Spanish—using human subjects. For English we used Amazon Mechanical Turk (AMT) with five annotators (known as *turkers*) per item. Using AMT provides annotations very quickly, possibly at the expense of reliability—we have restricted our task to turkers with at least 1000 approved annotations—, but it has been proven suitable for sense-disambiguation task (cf. Section 3.2.2). Given the demographics of turkers (cf. Ipeirotis (2010)), it is not possible to obtain annotations for every language using AMT. Thus, for Danish and Spanish, we obtained an-

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

notations from volunteers, most of them native or very proficient non-natives. See Table 4.1 for a summary of the annotation setup for each language. In addition to the data annotated by turkers, we have also obtained annotations by a single expert for the English data. We use the expert-annotated data to compare with the turker annotations.

Language	Annotators	Type
English	1	expert
	5	turker
Danish	3-4	volunteer
Spanish	6-7	volunteer

Table 4.1: Amount and type of annotators per instance for each language.

In Section 4.1 we list the criteria we define for our annotation scheme. Section 4.2 describes the corpora and words we sample for our study, along with the preprocessing steps to obtain the final examples to annotate. The next three sections describe three different annotation environments: in 4.3 we describe the *expert* annotation method based on paraphrase, in 4.4 we describe the annotation method that uses crowdsourcing in Amazon Mechanical Turk (the *turker* variant of our annotation) for English, and finally Section 4.5 describes the scheme followed by *volunteer* annotators for Danish and Spanish.

The expert scheme is only followed by one annotator (namely the author of this dissertation), but both the volunteer and the turker schemes generate data where each example receives more than one annotation. We need to determine one single sense tag from all the annotations an example might have. In 4.7 we describe the implications of assigning a sense tag from a series of annotations and propose two methods: VOTE, a majority-voting method with a theory-compliant backoff strategy and MACE, an unsupervised Expectation-Maximization method that scores annotators according to their estimated reliability. Finally, we compare these two methods in Section 4.8.

4.1 Design criteria for an annotation guideline

The goal of our annotation task is to collect semantic judgments on the meaning of examples of dot-type nominals. Each example is annotated individually following a token-based approach, instead of annotating the behavior of the overall dot type. Following the practices and caveats collected in Section 3.2, we establish the criteria for our annotation scheme:

1. Our method is **class-based**, and this requires the annotation scheme for regular polysemy to be consistent across the whole regular polysemous semantic class, which implies there is one sense inventory per semantic class (i.e. per dot type) and not per lemma. A class-wise sense inventory mitigates some of the complications of an annotation scheme for word senses, and also complies with regular polysemy being a class-wise phenomenon.

2. The size of the **sense inventory** has to be kept to a minimum in order to avoid the problems of an overcomplicated annotation scheme. Problems

like difficulty in annotation (cf. Section 3.2) or under-representation of some very fine-grained senses penalize agreement. In our case, we cannot do away with the *literal* and *metonymic* senses, and we also include an *underspecified* sense, for a total of three.

3. Our unit extent of annotation is the **word level** (cf. Section 2.5.3). We define an example to annotate as one sentence with one single *headword* that the annotators have to label as either *literal*, *metonymic* or *underspecified*.

4.2 Data

This section describes how the examples for all the datasets have been obtained from corpora by querying concordances of the lemmas of the dot types in Rumshisky et al. (2007).

For each of the nine dot types (five for English, two for Danish, two for Spanish) listed in Section 2.6, we have randomly obtained a series of corpus examples, preprocessed them, and selected the first ranked 500 that were not removed during preprocessing (cf. Section 4.2.1). The value of 500 examples was experimentally obtained in a previous study (cf. Appendix B).

Each example consists of a sentence with a selected *headword* belonging to the corresponding dot type. For English and Danish we used freely available reference corpora: the ANC for English (Ide and Macleod, 2001), and KorpusDk for Danish (Andersen et al., 2002). For Spanish we have used IulaCT, a corpus built from newswire and technical text (Vivaldi, 2009).

Even though we are studying data in three different languages, we have not used parallel corpora. Using translated texts would allow us to compare the datasets sentence-wise, but it would also incorporate calques from the source language which would undermine the validity of these comparisons.

For most of the English datasets we used the lemmas in Rumshisky (Rumshisky et al., 2007) also listed in Section 2.6, except for ENG:LOCORG. For Danish and Spanish we translated the lemmas from English and expanded the lists using each language's wordnet (Pedersen et al., 2009; Gonzalez-Agirre et al., 2012) as thesaurus. The Danish and Spanish corpora are smaller than the English corpus and this expansion of the list of lemmas was necessary to ensure that the datasets would reach 500 examples after the discarding that takes places during preprocessing.

For the three version of LOCORG we used high-frequency names of geopolitical locations—continents, countries and cities—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 across datasets: *Afghanistan, Africa, America, China, England, Europe, Germany*, and *London*. The list of words for each dataset is provided in Appendix A.1.

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* (cf. Section 2.6.1 for a more detailed explanation on preemption). Our claim is that this lexical phenomenon does not impede metonymy for animal names, it just makes it less likely. In order

to assess this, we have included 20 examples of *cow*[2]. The rest of the dataset consists of animal names that do not participate in this lexical alternation, like *eel*, *duck*, *chicken*, or *sardine*.

4.2.1 Preprocessing

After obtaining all the candidate examples for the annotation, we discarded examples that were shorter than five tokens, showed bad sentence splits (e.g. sentences interrupted after an unrecognized abbreviation), homonyms (e.g. a novel called *Paris*), or other kinds of polysemy like metalinguistic usages or object-for-representation metaphors (cf. Section 2.3.1). The discard rate can be very drastic; for instance, for ENG:CONTCONT we discarded about 800 examples in order to have 500 examples to annotate.

We want to minimize the bias of giving too much presence to a certain lemma in a dataset. For instance, building the whole dataset for ENG:LOCORG only from examples with *Moscow* as a headword would not make it representative of the overall dot type. To mitigate this bias, we used a sorting method to select which examples to incorporate into the 500 examples for each dataset.

We provide an illustration on the method chosen to select examples to compensate for the skewedness of word distribuions. Say we have three examples for *Paris*, two for *London*, two for *Rome*, and one for *Berlin*; and we want to select six of them. Firstly, we sort the preprocessed candidate examples randomly and give them a numeric index. This yields a list C of lists with sorted candidate examples. The lists within C are sorted alphabetically by the name of their lemma. We use a list called E to store the chosen examples for the dataset. The data structures have the following content upon initialization:

$$C = [[Berlin_1], [London_1, London_2], [Paris_1, Paris_2, Paris_3], [Rome_1, Rome_2]]$$
$$E = []$$

In each iteration i, we add the i-th example of each sublist to E and remove it from C. In the first iteration, we add all the examples with the index 1 to E and remove them from C. After this iteration C has four examples:

$$C = [[London_2], [Paris_2, Paris_3], [Rome_2]]$$
$$E = [Berlin_1, London_1, Paris_1, Rome_1]$$

In the second iteration there are no more examples left for *Berlin*, and we only have to choose two examples more to reach the desired amount of six. Following the alphabetical order, we choose examples with an index of 2 for *London* and *Paris*

$$C = [[Paris_3], [Rome_2]]$$
$$E = [Berlin_1, London_1, Paris_1, Rome_1, London_2, Paris_2]$$

At the end of the second iteration, E has reach the desired cardinality of six examples and there are two examples in C that are not incorporated into the dataset. Note that this method compensates for skewness because it maximizes the entropy of a certain lemma within a dataset, but it does not avoid it altogether. In this illustration, *Berlin* is a hapax and only appears once in the

[2]Indeed, two of the examples for *cow* were annotated as metonymic by turkers.

dataset. For some of the datasets that had to be expanded with more lemmas to make sure that E reached 500 examples, there are also hapaxes, like *anchovy* and *yak* in ENG:ANIMEAT. All hapaxes are included in the first iteration because they are all the first and only element of their sublist in C.

4.3 Expert annotation scheme

We have devised an expert annotation scheme to be able to compare against the results obtained by crowdsourcing in the turker scheme. Obtaining expert annotations also allowed us to get hands-on experience with the data and acquire intuition on what to expect in terms of sense selection and difficulty of examples. However, annotating 500 examples is cumbersome, and we only do it for English (for a total of 2500). For Danish and Spanish we are using volunteers as annotators, and we consider that there is no need to build a parallel annotation to evaluate them.

The annotation scheme for the expert uses a paraphrase test to determine the sense of each nominal headword h. In Table 4.2 we provide the paraphrases we established for each sense. For each sentence, the paraphrase scheme is as applied as follows:

1. If h can be paraphrased using the paraphrase for the literal sense, tag with a literal sense tag. ANIMAL, ARTIFACT, CONTAINER, LOCATION, PROCESS are literal.

2. If h can be paraphrased using the paraphrase for the metonymic sense, tag with a metonymic sense tag. MEAT, INFORMATION, CONTENT, ORGANIZATION, RESULT are metonymic.

3. If both paraphrases **can been applied** to h, or **must be applied** for h to make sense, consider h as underspecified.

Table 4.2 lists the paraphrases for each sense. Literal senses are listed first for each pair.

Sense	Paraphrases
LOCATION	"the place called X", "the territory of X"
ORGANIZATION	"the people or institutions of X"
ARTIFACT	"the physical object of X"
INFORMATION	"the content of X","what X says"
CONTAINER	"the X as such", "the container of X"
CONTENT	"the content of X","what X contains"
PROCESS	"the event of X", "the process of X"
RESULT	"the result of X happening", "the result of having done X"
ANIMAL	"the animal X"
MEAT	"the meat of X"

Table 4.2: Paraphrase listing for the different senses

Using paraphrases as a way to identify senses is not devoid of problems. During the annotation, some of the paraphrases, in particular for PROCRES, felt

rather ad hoc. Accepting a paraphrase as valid, i.e. as preserving the meaning of the original statement, becomes more difficult if the phrasing is awkward. We do not expect non-experts to be comfortable accepting an awkward paraphrase as semantically valid regardless of pragmatics, and we only use paraphrasing as a method to assign word senses for the expert scheme.

Using a paraphrase tests implies that we are working with the assumption that meaning is mostly preserved from a sentence to its expanded paraphrase. Vila Rigat et al. (2011) notice that the notion of preservation of meaning is controversial. Since our scheme consists in comparing between different paraphrases and choosing the most felicitous one, we are not bound by a strict adherence to a notion of absolute preservation of meaning.

Table 4.3 shows the sense distributions from the expert annotation broken down in literal (L), metonymic (M), and underspecified (U) sense. We compare the sense distributions of expert, turkers and volunteers in Section 4.6.

Dataset	L	M	U
ENG:ANIMEAT	0.65	0.21	0.15
ENG:ARTINFO	0.12	0.65	0.24
ENG:CONTCONT	0.68	0.15	0.17
ENG:LOCORG	0.54	0.22	0.24
ENG:PROCRES	0.36	0.43	0.22

Table 4.3: Sense distributions for expert in relative frequency

4.4 Turker annotation scheme

Expert annotation is both slow, costly and, in this particular scenario, biased. That is, an annotator that has read enough theory on the GL to think in terms of dot-type sense selection and underspecification will be biased in his understanding, and might be all too eager to over-recognize examples as underspecified. We therefore want to measure whether the non-expert annotation by several people converges around the underspecified senses the expert has annotated.

Furthermore, it is a standard practice to employ agreement measures to assess the validity of annotation schemes. Agreement measures require more than one annotation per item, and using only one expert is thus not sufficient to calculate agreement. To obtain more than one annotation from non-experts for each example, we use Amazon Mechanical Turk (AMT).

But using AMT has its downsides, as complex annotation schemes are more noise-prone. An annotation scheme for AMT needs to prime simplicity, and a paraphrase scheme would add an additional layer of complexity to the task for a non-expert user.

Indeed, instead of providing a sentence with a headword to disambiguate and the question "what can this headword be paraphrased as?", the scheme offers the sentence and the highlighted headword along with the more straightforward question "what does the headword mean?". Each annotation task for a given semantic type has an illustrating example of both the literal and metonymic senses. We did not provide examples of underspecified examples to avoid priming the annotators with our intuitions about underspecification.

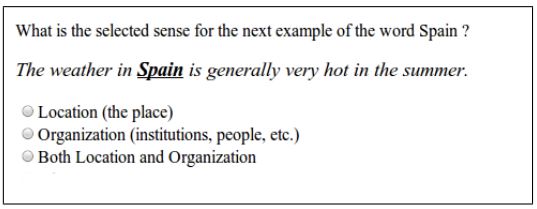

Figure 4.1: Screen capture for a Mechanical Turk annotation instance or HIT

Each dataset, a block of 500 sentences with a headword belonging to a dot type, is 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 4.1 shows an instance of the annotation process.

In Table 4.4 we find the glosses to describe each sense for AMT. Notice how the definitions are simplified and try to be more intuitive than the paraphrases in Table 4.2, which were only used by the expert.

Sense	Gloss
LOCATION	the place
ORGANIZATION	institutions, people, etc.
ARTIFACT	the physical object
INFORMATION	the information, the content
CONTAINER	the object itself
CONTENT	what is inside
PROCESS	the act of doing something or something happening, possibly placed in time
RESULT	the result or output of the action
ANIMAL	the animal itself
MEAT	its meat or flesh

Table 4.4: Sense glosses for turkers

In addition to its 500 corpus examples to annotate, each dataset had 5 synthetic examples that were annotated by 10 turkers. Our intention when including synthetic examples was to assess the feasibility of the whole system, because if the performance was low for simple, made-up examples where we had an expectation of what sense they should receive, it would not be possible to obtain conclusive data using AMT on actual corpus data. We considered this pilot study satisfactory and carried through the annotation task proper.

These additional synthetic examples and their annotations are described in Appendix A.3. These examples were only annotated by turkers for reference purposes abd were not included in the datasets used for the NLP experiments in the following chapters.

Dataset	L	M	U
ENG:ANIMEAT	0.70	0.26	0.04
ENG:ARTINFO	0.51	0.36	0.13
ENG:CONTCONT	0.64	0.28	0.08
ENG:LOCORG	0.59	0.35	0.06
ENG:PROCRES	0.36	0.56	0.08

Table 4.5: Raw sense distributions for turkers in relative frequency

Table 4.5 shows the proportion of annotations for each sense in each datasets out of all the judgments of all turkers. Before determining the final sense tag for each element, we can draw some expectations about the final sense distributions.

The underspecified sense is systematically less frequent for turkers than for the expert (cf. Table 4.3), but the ranking by frequency of the three senses is the same, with the exception of the ENG:ARTINFO dataset, where the expert annotated fewer (0.12) literal examples. However, the literal sense takes 0.51 of the turker annotations for this dataset. Notice that this is an estimate of the overall behavior of turkers before we assign definitive a sense tag to each sense. Cf. Section 4.7.1 and 4.7.2 for the resulting distributions after applying a sense-assignment method.

ENG:ANIMEAT is the dataset with fewest underspecified annotations and the highest preference for the literal sense. On the other hand, ENG:PROCRES is the dataset with fewest literal annotations, and turkers lean more towards a metonymic (RESULT) interpretation when analyzing word of this dot type. We provide a more detailed analysis of the commonalities and differences between types of annotators in Section 4.9.1

4.5 Volunteer annotation scheme

Using AMT is fast, easy and affordable. Still, it is difficult if not right down impossible to obtain annotations for languages with fewer speakers, like Danish. Even though Spanish datasets have been successfully annotated with AMT (Mellebeek et al., 2010; Irvine and Klementiev, 2010), we have tried to obtain volunteer annotators for our Danish and Spanish data.

For Danish and Spanish we have chosen to annotate the LOCORG and CONTCONT dot types. We split the 500 preprocessed examples of each dataset in blocks of 50 and distributed them amongst the annotators in spreadsheet files.

The annotators were given instructions like the ones shown in Figure 4.1 to assign one of the three possible senses to the headword of each example. Like in the turker annotation setup, we did not provide an example of the underspecified sense to avoid biasing the annotators.

A total of 46 annotators participated in the Danish task, yielding three or four annotations per example. For Spanish we obtained more volunteers (65) and each example was tagged by between six and eight annotators.

Dataset	L	M	U
DA:CONTCONT	0.65	0.20	0.15
DA:LOCORG	0.66	0.20	0.14
SPA:CONTCONT	0.44	0.34	0.22
SPA:LOCORG	0.45	0.35	0.20

Table 4.6: Raw sense distributions for volunteers in relative frequency

Table 4.6 shows the proportion of annotations for each sense in each datasets out of all the judgments of all volunteers. The distributions are similar between the two Danish and the two Spanish datasets. This indicates that volunteers have a similar bias.

4.6 Annotation results

In this section we cover the measures used to quantify the agreement of human-annotated datasets, and examine the agreement scores of our data.

4.6.1 Agreement measures

A measure of agreement is used to assess the reliability of an annotation task. Since we have no intrinsic way of knowing the validity of the annotated data unless we have a previous gold standard (or conduct an extrinsic evaluation using the data as training data), we use agreement metrics to describe the reliability of the annotation task as a proxy for the validity of the data.

Note that a task where all the annotators said exactly the opposite of what they should would have perfect agreement, but not would be valid at all. Nevertheless, such adversarial annotators are not expected, and agreement values are often interpreted as an indicator of the quality of the data.

The simplest agreement coefficient is observed agreement (A_o). Observed agreement gives the proportion of how many pairwise comparisons among the annotators have the same value.

However, the A_o coefficient has no chance correction, that is, there is no adjustment on the agreement value to account for how much agreement is expected by chance. For all non-perfect ($A_o < 1$) agreement values, chance-corrected agreement will be strictly lower than observed agreement, because some chance agreement is always to be expected.

Cohen's κ is an agreement coefficient adjusted for chance agreement. Originally defined for two annotators, Cohen's κ can be generalized to any number of annotators. Chance agreement is calculated under different assumptions depending on the metric. Cohen's κ assumes a different distribution of categories for each annotator to reflect annotator bias.

The κ coefficient systematically overestimates chance agreement when one value is more frequent. This results in scores that can be unfairly low, given that sense distributions are often skewed.

Krippendorf's α disregards annotator bias and calculates the expected agreement by looking at the overall distribution of judgment regardless of which coder is responsible for each specific judgment.

According to Artstein and Poesio (2008), Krippendorf's α is used in computational linguistics in order to assess the *reliability* of an annotation procedure, as it calculates agreement disregarding annotator bias. But κ is calculated taking individual bias into consideration in the chance agreement, which means that the κ value informs on the *validity* of a specific annotation task. However, for large amounts of data, α and κ tend to converge.

There is dissent between the acceptability of an annotation task given its agreement. Artstein and Poesio note that are agreement coefficients above 0.40 are considered good agreement in the medical literature. A value of 0.80 or higher in a chance-corrected coefficient is considered impeccable for many tasks, but they admit the difficulty of word-sense related annotation tasks. Following this thought, Passonneau et al. (2012) interpret values above 0.50 with unweighed α as good agreement in their evaluation of the sense-annotated MASC corpus. For other sense annotation tasks and their agreement values, cf. Section 3.2.2.

Each example in English is annotated by five annotators from AMT. However, the annotators vary from semantic class to semantic class, and even the total number of annotators is different in each annotation task. Therefore we describe the reliability of the data in terms of Krippendor's α, as it does not consider annotator bias in its calculation of chance agreement. Using α as a reliability measure for meaning-based annotation tasks has become a common practice (Zarcone and Padó, 2010; Passonneau et al., 2012), which we also adscribe to.

Note that even though the underspecified sense subsumes the two other senses, we have calculated agreement considering senses as mutually exclusive, without considering that the underspecified could be a double tag (literal and metonymic) or any variation on the usual sense independence assumption taken when calculating agreement for word sense annotations. Even though α can provide agreement for multiple tags or even ordinal annotations, we calculate α using sense independence to make the coefficients comparable to the state of the art (cf. Section 3.2.2).

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Table 4.7: Strength of agreement

Interpreting agreement scores

Understanding observed agreement is relevant for the experiment in Chapter 9 and to estimate the quality of our annotation procedure. Observed agreement A_o is interpretable as the proportion of matching items between annotations only if there are two annotators. For any higher amount of annotators, the value of A_o is not the proportion of matching items but the proportion of match-

ing pairwise comparisons out of the possible $\binom{|A|}{s}$, where $|A|$ is the number of annotators and s is the number of possible senses.

In this section we detail the values of A_o that five annotators (the amount of annotators in our AMT setup) can generate when annotating three possible senses. Given one single annotated item (a sentence with a headword to disambiguate), five annotators, and three possible senses (A, B, C), a distribution of categories for that item (D) ensues. The average observed agreement inter-encoder agreement (A_o) for one item has the next possible values in this scenario: 1.0, 0.6, 0.4, 0.3 and 0.2.

1. An A_o value of 1.0 means perfect agreement; all five annotators have picked the same category:

$$D = [A, A, A, A, A]$$

2. An A_o value of 0.6 indicates a majority voting of 80%; four out of five annotators have picked the same category, and only one disagrees:

$$D = [A, A, A, A, B]$$

3. An A_o value of 0.4 indicates a 60% majority; three have chosen category A and two have chosen category B:

$$D = [A, A, A, B, B]$$

4. An A_o value of 0.3 also indicates a 60% majority, but with lower agreement; three have chosen category A, and the other two have picked B and C respectively:

$$D = [A, A, A, B, C]$$

5. An A_o value of 0.2 indicates the possible lowest agreement, where the distribution of categories is as even as possible (and has therefore the highest entropy):

$$D = [A, A, B, B, C]$$

4.6.2 Agreement scores

The purpose of measuring inter-annotation agreement is to assess the reliablity of the annotation scheme. Coefficients like Krippendorff's α report the proportion of observed agreement that exceeds the agreement that would be expected if annotators labelled items at random, given some probability of each label.

As mentioned previously in this section, there is no consensus in the accepted value of κ or α that indicates that an annotation task has issued a result of enough quality. In a recent annotation evaluation, Passonneau et al. (2012) consider that an α of 0.50 is good agreement. Still, some words have senses that are easier to annotate, and while the noun *strike* and the adjective *high* yield α scores over 0.80, the adjective *normal* and the verb *ask* give α values of -0.02 and 0.05 respectively. In their analysis, they also note that items instances with low agreement showed unusual vocabulary or complicated syntactic structures.

We have mentioned the α coefficient is a measure of reliability of an annotation procedure. But if all the annotation-setup variables are the same, and the only thing that changes is the input data (the dot type and the particular corpus examples), we can determine that some dot types are easier to annotate than others, given that the easier ones will have higher scores, because the method is more reliable for them.

Our scheme has a sense inventory of the same granularity for each semantic class, but we have no control over the intrinsic difficulty of the senses in the dot type—e.g. is ANIMAL easier to grasp than PROCESS?—, nor over the difficulty of the contexts themselves. We expect variation in α score between all the datasets in this dissertation, as the sense differences are easier to grasp for some specific dot types. After the annotation task we obtained the agreement values shown in Table 4.8.

Dataset	$\overline{A_o} \pm \sigma$	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 4.8: Averaged observed agreement and its standard deviation and α

Average observed agreement ($\overline{A_o}$) is the mean across examples for the pairwise proportion of matching senses assigned by the annotators. Krippendorff's α 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 ($\alpha > 0.61$) for one dot type and *moderate* ($\alpha > 0.41$) for four. 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.

But also in the low-agreement datasets there are perfect-agreement examples—56 for ENG:ANIMEAT and 66 for ENG:PROCRES—, which indicates that the disagreement is not homogeneously distributed. We claim that variation in agreement is a result of a difficulty of annotating certain examples. We explore this clain in Chapter 9.

Again, note the value of α is only a proxy for a measure of the quality of the data. In Chapters 5, 7, 8, and 9 we train and test our system on all the datasets, thus evaluating their quality extrinsically, and determining to which extent the performance of the system is hampered by the variation in α score.

4.7 Sense assignment methods

In this section we describe to ways to assign a final sense tag from all the annotations a item has received from either turkers or volunteers. We describe two sense assignment methods (SAM), namely VOTE and MACE.

In section 4.6 we provide agreement values for the annotation task, but we have not yet determined which is the sense tag to assign to each example from the n tags that it has received from either turkers (in English) or volunteers (in Danish and Spanish). We need to find a method to determine a final sense tag for each example in each dataset. In each dataset, for each sentence with a dot-type headword h to disambiguate, there is a set of annotations available, namely:

1. For English, the judgment by an expert (the author of this dissertation). These judgments have been obtained with the paraphrase test described in 4.3.

2. For English, the annotations collected from Amazon Mechanical Turk. These judgments have been obtained using the annotation scheme described in 4.4.

3. For Danish and Spanish, the annotations collected from the volunteers, following the scheme described in 4.5.

However, there are many ways in which a final sense tag for each example can be derived from these judgments. We used two different SAMs to obtain a final sense tag assignment for the data, only using the annotations provided by either turkers or volunteers, without incorporating the expert annotations in our gold standard.

One of the reasons not to incorporate the expert annotations in the final gold standard because they have no agreement score themselves. Even though we could provide a chance-corrected agreement score like α between the expert and the turker annotations, this value would not be informative of the reproducibility of the annotation task in a straightforward fashion, mainly because the way that the sense judgments have been obtained in each scheme is different.

In this way, even though the expert annotation is likely more internally consistent—a single annotator with the same bias for five datasets—, we only used it to compare the turker annotation and not for machine learning experiments. We claim that it is more realistic to use crowdsourced annotations to generate our gold standard than the annotations of one expert with a theoretical bias.

Nevertheless, we use the expert annotation as gold standard to compare the output of the two sense assignment methods in Section 4.8.

4.7.1 Voting method

In this section we describe VOTE, a SAM based on majority voting with a theory-compliant backoff strategy to resolve ties. Using majority voting to determine the sense tag is a fairly straightforward method that does not seem to commit too much to a certain theory of meaning: one determines the final sense tag from the sense that has received the most annotations for a given example and

has thus *plurality* (or *simple majority*). However, a problem appears when there are ties between candidate senses and there needs to be chosen one.

If using majority voting, resolving ties by assigning the literal (or most frequent) sense is not acceptable for every tie. E.g. if five annotators annotate as example twice as literal, twice as literal and once as underspecified ($A = [U, U, M, M, L]$), how can we justify the customary backoff to literal ?

It is commonplace to backoff to the most frequent sense in case of disambiguation of homonymy or non-regular polysemy. In the case of regular polysemy, Nissim and Markert (2005a) backoff to the literal sense when they find ties. In their data, their literal sense is also the most frequent one. A glance at Table 4.3 reveals that, at least for the expert, two of the datasets are skewed towards the metonymic reading. When describing the ARTIFACT•INFORMATION in Section 2.6, we took the stance of considering ARTIFACT as the literal sense, even though the INFORMATION sense appears in about 65% of the examples in ENG:ARTINFO.

This skewness towards INFORMATION could be an argument to actually consider the INFORMATION sense the literal one. But in the ENG:PROCRES we also see that RESULT is more frequent than PROCESS, and we have no doubts about PROCESS being the literal sense in this dot type. We have also mentioned in Section 3.2.2 that the literal sense need not be the most frequent one; a corpus with a lot of recipes would have the metonymic sense MEAT as most frequent for the ANIMAL•MEAT dot type.

For these reasons, we consider that backing off to the literal sense is not justifiable, especially when the tie is between the metonymic and the underspecified sense. Backing off to the most frequent sense is not better either. Moreover, we conceptualize the underspecified sense to be placed between the literal and metonymic, and consider it a more proper candidate to back off to when there are ties between literal and metonymic senses.

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. In the VOTE SAM, ties are resolved as follows:

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**.

This way of resolving ties aims to reproduce the logic followed when using the expert annotation scheme by placing the underspecified sense between the literal and the metonymic. We take it as a representation of Dirven's (cf. Section 2.3) understanding of the literal-to-figurative continuum. This is however a scale based on an author's intuition (and our compliance) and has no assessed psycholinguistic validity.

This SAM's major advantage is that, instead of depending on the frequency of senses, it resolves the sense tag locally for each example and can be applied to any number of annotations and examples, no matter how few.

Table 4.9 provides the relative frequencies for the senses of all datasets after appling VOTE.

Dot type	L	M	U
ENG:ANIMEAT	0.72	0.27	0.01
ENG:ARTINFO	0.28	0.61	0.11
ENG:CONTCONT	0.71	0.24	0.05
ENG:LOCORG	0.61	0.34	0.04
ENG:PROCRES	0.31	0.60	0.09
DA:CONTCONT	0.65	0.16	0.18
DA:LOCORG	0.64	0.19	0.17
SPA:CONTCONT	0.58	0.28	0.14
SPA:LOCORG	0.63	0.28	0.09

Table 4.9: Literal, Metonymic and Underspecified sense distributions over all datasets for VOTE

Notice how VOTE gives the metonymic sense the most frequent one for ENG:ARTINFO, even though the literal sense had 0.51 of the probability mass of the raw annotations (cf. Table 4.5). The sense distribution with VOTE is thus more similar to the one provided by the expert. This change in the resulting sense distribution is a consequence of the homogenous distribution of the annotations for the literal sense in this dataset, which get filtered out by the combination of plurality-based sense assignment and our backoff strategy .

In Table 4.10 we break down the amount of underspecified senses of each dataset into those obtained by simple majority (i.e. plurality) or backoff. We give the values in absolute frequency to ease the comparison with the additional columns. The columns labelled L, M and U provide the sense distributions for each dot type. We provide the sense distributions in absolute frequencies to make the comparisons more immediate between this method and the method described in 4.7.2.

Dataset	L	M	U	P	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
SPA:LOCORG	314	139	47	40	7

Table 4.10: Literal, Metonymic and Underspecified sense distributions over all datasets, and underspecified senses broken down in Plurality and Backoff

The preference for the underspecified sense varies greatly, from the very infrequent for ENG:ANIMEAT 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 P and B show respectively whether the underspecified senses are a result of plurality in voting (the underspecified sense being the most frequent for a certain item) or backoff (a tie between the literal and the metonymic).

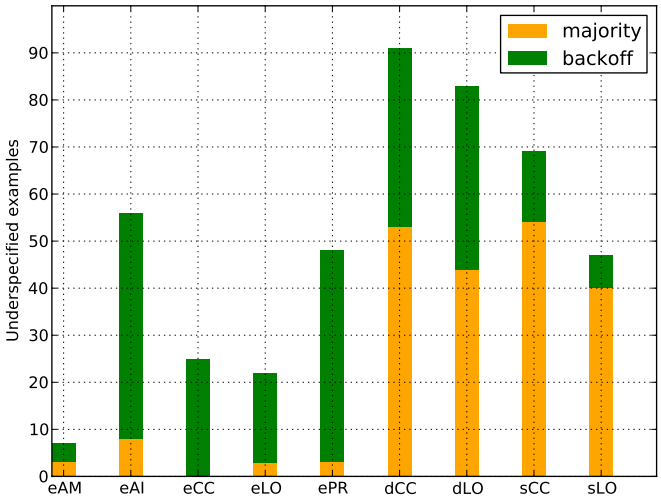

Figure 4.2: Underspecified senses with VOTE, with Plurality and Backoff

Figure 4.2 illustrates the P and B columns from 4.10. Each bar shows the amount of underspecified sense, with the amount of underspecified senses that fall under P (underspecified senses assigned by plurality) in yellow, and the underspecified senses assigned by backoff (B) in green. The names of the bars are abbreviations for the dataset names, in the same order they are displayed in Table 4.12, starting with ENG:ANIMEAT (eAM) and finishing with SPA:LOCORG (sLO).

In contrast to volunteers, turkers disprefer the underspecified option, and most of the English underspecified sense tags are assigned by backoff when there is a tie between the literal and the metonymic. Note that ties between two senses when there are five annotators are only possible if the third sense that is not involved in the tie appears once.

Nevertheless, 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 datasets (ENG:ARTINFO, ENG:PROCRES) show majority of metonymic senses, which are always second in the scheme (cf. Fig. 4.1). These two datasets also have majority of metonymic senses in the expert annotation. We expand the remarks on annotation behavior in Section 4.9.

4.7.2 MACE

The data for each language has been annotated by a different amount of people, either volunteers or turkers for each different dataset. Both the number and the kind of annotators has an impact on the result data. More reliable annotators

are expected to generate more reliable data, and a higher number of them decreases chance agreement. Besides using majority voting with the backoff strategy detailed in the previous section, we use MACE (Hovy et al., 2013), an unsupervised system, to obtain the sense tag for each example.

We use MACE as a SAM because we want to compare our voting sense assignment method with a method that is not informed of our theoretical assumptions about the underspecified sense as a sense between literal and metonymic, nor that does not simply backoff to the most frequent sense in case of ties.

MACE (Multi-Annotator Competence Estimator) is an unsupervised system that uses Expectation-Maximization (EM) to estimate the competence of annotators and recover the most likely answer by adjusting the weights of the annotators. 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).

While removing annotators improves the agreement, using MACE, we can obtain weights for the different annotators instead of simply removing those that show less agreement. The method is aimed at identifying spamming behaviors like the one shown by annotators that always click on the same value, or annotators that annotate at random. MACE can be tuned for a confidence threshold to only resolve labels with a certain confidence. While doing this could be interesting as an indirect means of identifying underspecified senses (some of which would have low confidence), we run MACE on its default settings: 50 iterations, 10 random restarts, a smoothing coefficient of 0.01 and no usage of the confidence threshold option. For a more detailed explanation of MACE, refer to Hovy et al. (2013).

We have trained MACE individually for each dataset to avoid introducing noise caused by the difference in sense distributions, which would alter the likelihood of each sense tag. We claim that datasets that show less variation between senses calculated using majority voting and using MACE will be more reliable.

Table 4.11 provides the relative frequencies for the senses of all datasets after appling MACE.

Dataset	L	M	U
ENG:ANIMEAT	0.68	0.29	0.03
ENG:ARTINFO	0.34	0.36	0.30
ENG:CONTCONT	0.59	0.35	0.06
ENG:LOCORG	0.58	0.39	0.03
ENG:PROCRES	0.31	0.42	0.27
DA:CONTCONT	0.45	0.27	0.28
DA:LOCORG	0.50	0.29	0.21
SPA:CONTCONT	0.54	0.31	0.15
SPA:LOCORG	0.60	0.29	0.10

Table 4.11: Sense distributions calculated with MACE

Along the sense distribution in the first three columns, Table 4.12 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 4.10 and 4.11. Table 4.12 provides the sense distributions in absolute frequency for ease of comparison with the two right columns.

Dataset	L	M	U	D	I
ENG:ANIMEAT	340	146	14	5%	3
ENG:ARTINFO	170	180	150	90%	46
ENG:CONTCONT	295	176	28	17%	0
ENG:LOCORG	291	193	16	8%	3
ENG:PROCRES	155	210	134	27%	33
DA:CONTCONT	223	134	143	24%	79
DA:LOCORG	251	144	105	21%	53
SPA:CONTCONT	270	155	75	7%	56
SPA:LOCORG	302	146	52	4%	40

Table 4.12: Sense distributions calculated with MACE, plus Difference and Intersection of underspecified senses between methods

Table 4.12 shows a smoother distribution of senses than Table 4.10, 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 20.6% and 29.6%.

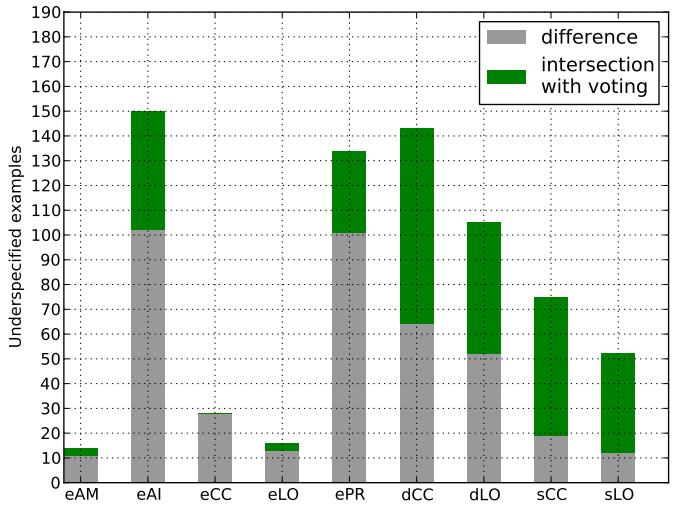

Figure 4.3: Underspecified senses with MACE, with Difference and Intersection

Figure 4.3 illustrates the D and I columns from 4.12. Each bar shows the amount of underspecified sense, with the amount of underspecified senses that fall under I (intersection between VOTE and MACE) highlighted in green. The

names of the bars are abbreviations for the dataset names, in the same order they are displayed in Table 4.12, starting with ENG:ANIMEAT (eAM) and finishing with SPA:LOCORG (sLO).

Although MACE increases the number of underspecified senses for all datasets but one (ENG:LOCORG), the underspecified examples in Table 4.10 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 VOTE, 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.

Assigning sense tags with MACE has the advantage of not being committed to a certain understanding of the relation between senses, something that is ingrained in the backoff strategy for VOTE. However, clustering methods also have biases of their own. Even though MACE runs EM a number of times with random restarts to overcome the local optima that some clustering algorithms can be prone to, it does seem to have a bias for evenly-sized clusters, something that becomes apparent in the schemes with fewer annotators or with lower agreement.

This tendency for evenness is an expectable shortcoming of a clustering method, because MACE needs a certain volume of data to converge consistently and it will update more drastically when applied to data with fewer annotators or lower agreement. Hovy et al have successfully applied MACE to the WSD datasets used in Snow et al. (2008), which have also been annotated by turkers. However, we know that word senses tend towards skewed distributions, which might complicate the applicability of MACE in our particular scenario.

4.8 Comparison between SAMs

For each of the datasets, we have two alternative SAMs, namely the one obtained by VOTE and the one obtained using MACE. In this section we provide a comparison between the voting method and MACE by comparing them to the expert-annotated data for English, and by comparing between themselves for the three languages.

4.8.1 Expert annotations versus raw AMT data

In Section 4.3 we have described the annotation scheme for experts and provided the sense distributions we obtained for English. In this section we provide a qualitative analysis on how the expert annotations map to the turker annotations, giving examples of agreement and disagreement between both kinds of annotators.

Before examining either SAM, we compare the raw annotations of the turkers with the expert annotations. Firstly we look at examples where the expert and the five turkers agreed fully on the literal and metonymic senses.

Perfect agreement examples

In example 4.1 we provide five examples, one for each English dot type, where the expert and the five turkers coincide in tagging the headword as literal.

(4.1) a) Atlantic *cod* in Canadian waters suffered a total population collapse and are now on Canada's endangered species.

b) In fact many people use *dictionaries* to press flowers, hide money or prop up uneven table legs.

c) And would a corkscrew and proper *glasses* be pushing it?

d) Visitors need a visa to enter *China*, and must disembark at the border (now called boundary) checkpoint, Lo Wai.

e) Another highlight of the last five genturies of geometry was the *invention* of non-Eucledian geometries (1823-1830).

These examples are obtained from the ANC and, even though they have perfect agreement in their annotations, they are less immediate in their interpretation than the examples we have provided in our theoretical framework Chapter 2, which were mostly synthetic.

Example 4.1.c), for instance, is a question that was shown to the annotators without any other context, and even though the situation is unclear—someone being at restaurant or being a guest at someone's, maybe it is even a rhetorical question by someone complaining about bad service—, it was clear to the annotators that the *glasses* meant in the example are containers.

The PROCESS•RESULT example, was tagged as literal by the turkers arguably because it is placed in a context where it is portrayed as something happening and delimited in time as an event. Example 4.1.b), is a fairly transparent example where the sentence suggests a series of alternative telic roles for *dictionary* that only take the ARTIFACT sense in consideration, disregarding the INFORMATION sense. These two datasets have low α (0.10 and 0.12), yet some of their examples also have perfect agreement.

In example 4.2 we provide five examples with the same words as in 4.1 that show perfect agreement and are unanimously tagged as metonymic.

(4.2) a) Imported dried *cod* (bacalhao) is the Portuguese national dish; several varieties are available, usually baked.

b) If meaning appears neutral, as it does in the modern *dictionaries*, it is only so following the dictates of objectivity, an ideology in itself, and the politics of invisibility.

c) Their vices are few—they enjoy an occasional *glass* of wine or beer, and now and then they may overindulge in chokolate-spiked trail mix.

d) *China* will prosecute the leaders of Falun Gong.

e) Among the rooms devoted to the history of science and technology, one gallery is reserved for Leonardo's *inventions*, displayed as models constructed from his notebooks.

Again, we also see the relative complexity of these examples. In these metonymic examples we find canonical cues for their interpretation. The example for *glass*, (4.2.c), could indeed be analyzed by the coercion view we have

not enforced in our analysis. In this case, a glass is enjoyed by *drinking* it, the telic role that selects for the CONTENT sense. With regards to the mass/count alternations, in the example for *cod*, the animal is used as a mass noun and is referred to as a dish. The metonymic sense RESULT is first and foremost identifiable by the word *inventions* being in plural and thus being a count noun, but also being headed by the genitive of the invention's author's name.

Expert-mechanical turk mistmaches

Unfortunately, perfect-agreement examples are only 856 out of 2500 (34%) for English. In the five English datasets there are also 203 examples where the expert annotation is not present in any of the five annotations by turkers, and a SAM that applies weights to annotators would not be able to reproduce the expert judgment. Almost all (96%) of these mismatching examples were tagged as underspecified by the expert. Example 4.3 shows one example from each dataset where the expert annotation is not included in the example's five turker sense annotations.

(4.3)
a) In a region without any noteworthy industry other than *cod* fishing, little forestry, and tourism, the people add to their income by selling their craftwork and farm produce from improvised stands.
b) Both *dictionaries* list abbreviations and biographical and geographical entries in separate sections at the back, a practice I have never liked.
c) On the other hand, if the Canadian government wants to force Coca-Cola to print time-lapse photos of a nail dissolving in a *glass* of their product, then go for it.
d) Such practices are totally forbidden in *China*
e) On the contrary, respect for the capacity of the materials always wins out over daring visual *invention*, but a staggering technical imagination has also been summoned.

Most of the 203 examples in this group without overlap between expert and turker annotations show the highest possible disagreement between turkers for two possible sense tags (A_o=0.4), since there is a third sense that is only annotated by the expert only example (4.3.d) is an oddity because it has perfect agreement by turkers, who consider it a literal example. However, it accepted, according to the expert, both the organization and the location paraphrases. It seems the preposition *in* is a strong cue for a LOCATION reading, which turkers deem sufficient to assign the literal sense.

The rest of the examples were tagged two or three times as literal, and three or two times as metonymic by turkers. However, these examples were considered underspecified by the expert. In the example for *cod*, the expert has given the underspecified sense because in this sentence fish are both treated as animals and as the product of fishing. The example for *dictionary* has both its INFORMATION ("list abbreviations") and its ARTIFACT sense ("at the back") predicated.

4.8.2 VOTE and MACE against expert

The SAMs VOTE 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 VOTE and by MACE:

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

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

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

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

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

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

g) 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 VOTE or MACE, but not by both. Table 4.13 breaks down the five annotations that each example received by turkers in literal, metonymic and underspecified senses. The last three columns show the sense tag provided by VOTE or MACE from the turker annotations, and by the expert.

Example	L	M	U	VOTE	MACE	expert
a)	2	2	1	U	L	U
b)	3	1	1	L	U	L
c)	1	2	2	M	U	U
d)	2	2	1	U	M	M
e)	2	2	1	U	M	U
f)	3	0	2	L	U	U
g)	1	2	2	M	U	U

Table 4.13: 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 examples that are not clear-cut. The case for MACE is strong: in f), we consider the underspecified sense more adequate than the literal one obtained by VOTE, just like we are also more prone to prefer the underspecified meaning in g), 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 b) are empty and have no content yet, thereby being literal containers and not their content.

Examples a), d) and e) 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.

Note how example a) has the underspecified sense assigned by backoff in VOTE. This means that VOTE has chosen the least frequent annotation as sense label, because this scheme interprets the underspecified sense as sitting between the literal and metonymic.

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

These examples are useful to illustrate the different shortcomings of both SAMs, but we need quantitative measures to actually determine which is most adequate as a reference SAM out of VOTE or MACE. One way of doing it is by measuring accuracy, considering the sense assignment method as the predicted value and the expert annotation as the gold standard, or using an agreement coefficient between a SAM and and the expert.

Note that by comparing the output of the SAMs to the expert annotations, we are evaluating the SAMs in terms of which one best approximates the expert judgments, which we decide not to use as gold standard in the rest of our experiments.

However, we claim that using the expert annotation as gold standard for SAM comparison is a sound decision because all datasets have been annotated by the same expert with the same bias, and the judgments will be consistent across examples and across datasets. By comparing against the expert annotations, we have determining the consistency of the output of the SAMs.

In Table 4.14 we see that all the accuracies are higher for VOTE with regards to the expert annotation.

Dataset	MACE	VOTE
ENG:ANIMEAT	0.84	0.85
ENG:ARTINFO	0.50	0.64
ENG:CONTCONT	0.70	0.80
ENG:LOCORG	0.73	0.75
ENG:PROCRES	0.58	0.60

Table 4.14: Accuracy for VOTE and MACE with expert annotation as a gold standard

With regards to comparing the SAMs with the expert annotations using accuracy or an agreement coefficient, accuracy has the same value as measuring A_o for the two-annotator case between the expert and one of the predictions. Its informativeness is also limited. Since we are comparing one candidate to gold standard to an approximation, it is more informative se to build contingency tables in a manner similar to de Melo et al. (2012) and Boleda et al. (2008), and generate precision, recall and F1 measures for each sense and representation, instead of just using an aggregate measure where details are blurred.

Higher accuracies for VOTE mean that this sense assignment method does indeed fulfil its intention of approximating the expert annotation. But all the accuracies are higher for VOTE in Table 4.14, and the picture is very different when we look at the sense-wise F1 measure for each English dataset and sense assignment.

	MACE			VOTE		
Dataset	L	M	U	L	M	U
ENG:ANIMEAT	0.94	0.80	**0.21**	0.94	0.33	0.17
ENG:ARTINFO	0.41	0.62	**0.32**	0.50	0.81	0.21
ENG:CONTCONT	0.85	0.55	0.17	0.90	0.73	**0.27**
ENG:LOCORG	0.88	0.69	0.15	0.88	0.75	**0.21**
ENG:PROCRES	0.60	0.67	**0.38**	0.60	0.71	0.25

Table 4.15: F1 score over expert

Table 4.15 extends the comparative of sense distributions provided in Table 4.12. The VOTE SAM has a strong bias for the most frequent sense in any dataset, and MACE relaxes this behavior. If we look at the behavior of the underspecified sense, we see that no method is immediately best for capturing it in a satisfactory manner, or at least in a manner that resembles the expert's.

First of all, F-scores for the underspecified sense are fairly low. And even though the differences can seem drastic (0.32 to 0.21 for ENG:ARTINFO), their conclusiveness, seen either as differences in F-score or error reduction values, is dubious. However, VOTE seems to fare best with high-agreement datasets, which is also understandable because it aims to mimic expert annotation. But the underspecified in ENG:ANIMEAT are best captured by MACE, and this is a dataset with agreement on par with ENG:LOCORG.

The key difference between ENG:ANIMEAT and ENG:LOCORG is the sense distribution. In the expert annotation there are 72 examples in ENG:ANIMEAT tagged as underspecified. We have seen that MACE marks 14 examples as underspecified, and only three of them are also recognized by VOTE. In general, rule-based methods like VOTE will have higher precision, whereas statistical methods are more likely to provide better recall. This difference in performance seems to hold if we examine the precision, recall and F1-score values for the underspecified sense in all five English datasets.

	MACE			VOTE		
Dataset	p	r	F1	p	r	F1
ENG:ANIMEAT	0.64	0.12	0.21	1.00	0.10	0.17
ENG:ARTINFO	0.29	0.36	0.32	0.33	0.15	0.21
ENG:CONTCONT	0.36	0.11	0.17	0.60	0.17	0.27
ENG:LOCORG	0.62	0.08	0.15	0.68	0.12	0.21
ENG:PROCRES	0.34	0.42	0.38	0.42	0.18	0.25

Table 4.16: Performace of underspecified sense for MACE and VOTE using expert annotations as gold standard

In Table 4.16 we can see that, while VOTE always has higher precision when assigning the underspecified sense, it can suffer from lower recall, in particular in the low-agreement datasets ENG:ARTINFO and ENG:PROCRES, where ties are more frequent and less reliable. The ENG:ANIMEAT is, complementarily, too rigid and too stable in its annotations for ties to appear, and benefits from being resolved using MACE. Still, 14 underspecified examples resolved by MACE are a small subset of the 73 expected in the—maybe biased—expert annotation.

[say something here].we have an evaluation of the two different SAMs and how they relate to the expert annotation for English. We obtain better accuracy with VOTE, but if we break it down to sense-wise performance and focus on the underspecified sense, the difference in behavior becomes less evident.

While VOTE provides higher precision for the three high-agreement datasets, it does worse for the more difficult ENG:ARTINFO and ENG:PROCRES. A possible option would be to choose VOTE for high-agreement data to approximate the expert annotation and MACE for low-agreement data to improve the performance of the data when used for training or testing. But an agreement threshold should be then identified, and there is no guarantee that it would hold for the other two languages for which we have no expert annotation as a reference.

After these considerations, we settle for using VOTE as a reference SAM. First, the sense distributions yielded by MACE have an even-size bias which is typical of EM methods, while VOTE keeps the expectable skewness of senses. Second, VOTE is systematically more consistent with the expert annotation. Even though the expert annotation might overestimate the presence of the underspecified sense, VOTE is in general better at capturing the expert's judgments on the literal and metonymic senses.

4.9 Behavior of annotators

In this section we describe the behavior of turkers and volunteers with regards to their biases, and their possible dispreference for the underspecified sense. We complement this section with examples from the Danish and Spanish datasets to complement the comparisons for English in Section 4.2.

4.9.1 On the behavior of turkers

In Section 4.7 we see that turkers disprefer the underspecified sense. We hypothesize that, since turkers are not always native speakers, might lack nuances in their interpretation. But on the other hand, in NLP it is accepted that fluent non-natives can annotate data for English, as in Biemann and Giesbrecht (2011) or Markert and Nissim (2009). We have worked with this assumption ourselves when carrying out the expert annotation task with a fluent non-native as annotator.

Example 4.3.d) shows a sentence where the expert had predicted the underspecified example because something being "forbidden in China" issued a LOCATION and ORGANIZATION reading. However, the five turkers unanimously assigned the literal sense to this example. This is an indication that turkers might have a less nuanced understanding of this phenomenon, or that they focus on very obvious features like the preposition *in*.

However, a less-nuanced understanding of the task is not the only possible explanation for the turker dispreference for the underspecified sense. Turkers are also wary of their data being rejected for invalidity and might choose options with lesser perceived risk of rejection. Indeed, two of the turkers contacted us during the annotation task to inquire about their data being approved using majority rules, as they were concerned that discarding low-agreement turkers would be unfair on such a difficult task. It was explained to them that it was not the case.

It is also a possibility that turkers choose, in case of doubt, the easiest sense. The easiest sense is not necessarily the most literal one, and we already have seen that we do not find a general tendency to abuse first-option clicking, which would have caused a stronger bias for the literal sense. Endriss and Fernández (2013) provide a related account on the biases of turkers when annotating structured data.

We propose that turkers manifest a behavior that makes them choose the easiest option as a default in hopes of getting paid and not getting their annotations rejected. Whatever the case, turkers behave different than the expert and show a bias against the underspecified sense that is not as strong in data annotated by volunteers.

4.9.2 On the behavior of volunteers

The datasets for the dot types LOCATION•ORGANIZATION and CONTAINER•CONTENT have been annotated by volunteers for Danish and Spanish. For these datasets we do not have an expert-annotation to compare against, like we do for English.

Still, we can still contrast the general behavior of all annotators for all languages for these datasets. In Table 4.17 we show the raw distributions of senses chosen by the annotators before any sense assignment method was applied to the data. This is the overall proportion of times any annotator has marked an item as literal, metonymic or underspecified. We provide the LOCORG and CONTCONT datasets for all three languages. For English we provide the distribution from the turker annotation and from the expert annotation. Figure 4.4 reprents the information graphically.

Dot type	L	M	U
ENG:CONTCONT:EXP	0.68	0.15	0.17
ENG:LOCORG:EXP	0.54	0.22	0.24
ENG:CONTCONT:TURK	0.64	0.28	0.08
ENG:LOCORG:TURK	0.59	0.35	0.06
DA:CONTCONT	0.65	0.20	0.16
DA:LOCORG	0.65	0.21	0.14
SPA:CONTCONT	0.56	0.26	0.17
SPA:LOCORG	0.58	0.27	0.15

Table 4.17: Expert, turk and volunter sense distributions for the CONTCONT and LOCORG datasets

We can see that, for these two dot types, the literal sense is the most frequently chosen, regardless of language and type of annotator. For the under-

specified sense, however, we observe two particularities: the volunteer datasets have a proportion of underspecified senses that is consistent even across Danish and Spanish, and is more similar to the English expert datasets. Also, the ENG:LOCORG:EXP dataset stands out as the dataset where there is the highest proportion of underspecified sense tags being assigned, four times as often as in its turker counterpart ENG:LOCORG:TURK.

We are naively comparing proportions to obtain a qualitative assessment of the behavior of the annotators. When comparing the proportion of sense tags given by the expert to the proportion of sense tags given by the turkers or volunteers we are comparing the distribution over 500 sense tags against a distribution between 1500 and 3200 sense tags. Conducting independence testing would only make sense between annotations of the same data, and for Danish and Spanish we have no alternative to the volunteer to compare against.

Nevertheless, we can suggest that the behavior of volunteers sets the standard for what to consider the output of an annotation task by well-meaning (i.e. non-spamming), native annotators. Turkers share the same kind of bias for the most frequent, literal sense, but are less willing to give the underspecified tag for the reasons we suggested in the previous section, while the expert can be overzealous in his willingness to interpret a certain usage of a dot type as underspecified.

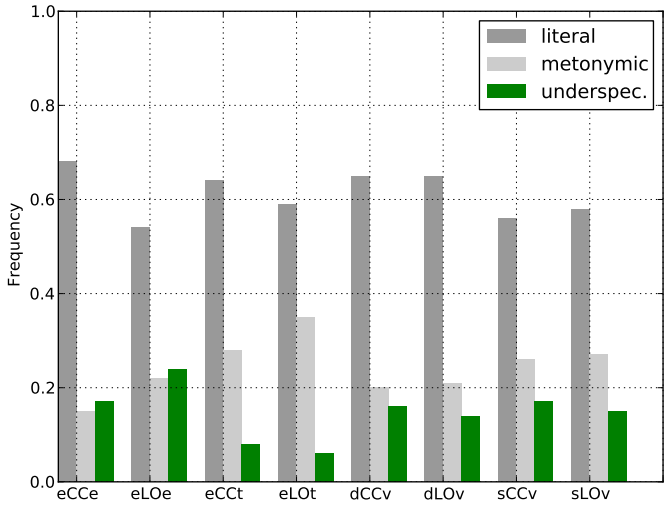

Figure 4.4: Distribution of annotations between the three senses for expert, turker and volunter for the CONTCONT and LOCORG datasets

Figure 4.4 illustrates the sense distributions in terms of relative frequency for the CONTCONT and LOCORG datasets. The first letter in the label of each three-bad group is the initial of the language (English, Danish, Spanish), the two middle letters stand for the dot type (CC for CONTCONT and LO LOCORG), and the final letter stands for the origin of the sense tags (**e**xpert, **t**urker, **v**olunteer). Thus, sCCV stands for Spanish-CONTCONT-volunteer. The datasets are pro-

vided in the same order as in Table 4.17.

Note how the proportion of underspecified sense is more internally consistent for datasets with the same kind of annotators, with the expert having the strongest bias for adjudicating the underspecified sense, turkers the least preference for doing so, and volunteers somewhat in the middle, regardless of language.

4.9.3 Danish and Spanish examples

In this section we provide examples of annotations with perfect agreement for the literal and metonymic sense for the Danish and Spanis datasets, examples that receive the underspecified sense by plurality (UNDR-PLURALITY) because the underspecified sense receives the most votes, and low-agreement examples that receive the underspecified sense by the backoff in Section 4.7.1 (UNDR-BACKOFF).

For the volunteer-annotated datasets there is no expert annotation. Instead, we provide an array of examples as we have done at the beginning of this section for the tuker- and expert-annotated English data.

We provide English translations within parentheses for the Danish and Spanish examples. The translations for the Danish examples show the definite form of a Danish noun as a hyphenated sequence of article and noun (e.g. *posen* becomes *the-bag*, cf. Section 2.6.6). In the Spanish examples we hyphenate the elliptic subject with the verb in the English translation (e.g. *tomé* becomes *I-took*).

For DA:CONTCONT, it seems that both examples that receive the underspecified sense have the CONTAINER and CONTENT sense active. In example c), the contents of the bag (*posen*) is what make the bag important and worth grabbing, and in d), *pot* (*kande*) refers to the coffee pot and what it could have contained, if it had indeed been served.

(4.5) a) LITERAL: Dolly stod nøjagtig hvor han havde forladt hende, hun holdt *skrinet* i hænderne og stirrede på ham.
(Dolly stood exactly where he had left her, she held *the-box* in the-hands and stared at him)

b) METONYMIC:For tiden er det globale forbrug 76,5 millioner *tønder* i døgnet.
(At the present time the global usage is 76.5 million *barrels* a day.)

c) UNDR-PLURALITY:Det har egentlig ikke generet os, for vi greb posen og gik med den, færdig.
(Actually it hasn't bothered us, because we grabbed the *bag*, and left with it, done.)

d) UNDR-BACKOFF: Kaffen kom, tre kopper—ikke en hel *kande*—og så oplevede de noget.
(The-coffe came, three cups—not a whole *pot*—and then they experienced something.)

For DA:LOCORG, we find that c) has a very limited context, and it is difficult to know what it means. Nevertheless, it has perfect agreement as an underspecified example. Likewise, if Berlin reminds of an asylum, it is likely that is

is because some human behavior which is geographically determined, thus the underspecified reading obtained by backoff.

(4.6) a) LITERAL: De norske ejere købte hesten på auktion i *England* for cirka 200.000 kr.
(The Norwegian owners bought the-horse in an auction in *England* for around 200,000 crowns.)

b) METONYMIC: Forholdet mellem Indien og *Kina* er ikke meget bedre.
(The-relation between India and *China* is not much better)

c) UNDR-PLURALITY: Og det er *Danmark*.
(And it is *Denmark*.)

d) UNDR-BACKOFF: Hvad indbyggerne angår, minder *Berlin* forlængst om en sindssygeanstalt.
(With regards to the residents, *Berlin* has resembled a mental asylum for a long time.)

For SPA:CONTCONT, it is the content of the sacks in c) what gives them the double reading as both containers and contents, much like the oil content of the cans in example d).

(4.7) a) LITERAL: Las páginas de Apple Computer, en cambio, enseñan entre otras cosas a construir una antena detectora de redes inalámbricas a partir de una *lata* de café.
(The pages of Apple Computer, on the other hand, teach among other things how to build a wireless network detector antenna from a coffee *can*.)

b) METONYMIC: Me tomé un *bote* de pastillas de mi madre.
(I-took a *bottle* of pills of my mother.)

c) UNDR-PLURALITY: Desde entonces , los *sacos* de yeso y cemento descansan sobre el suelo de su lujosa vivienda recién adquirida [...]
(Since then, the *sacks* of plaster and cement rest on the floor of the newly purchased, luxurious tenement [...])

d) UNDR-BACKOFF: [...] ha pedido a los ciudadanos que no compren ni una *lata* de aceite en los establecimientos de la firma anglo-holandesa.
(He-has asked the citizens not to buy a single oil *can* in any of the British-Dutch company's establishments.)

For SPA:LOCORG, example c) is underspecified with full agreement, arguably because the mentioned Spaniard has not been unable to adapt to the English weather or customs, or both. In example d) the annotators achieved mininum agreement on whether *Germany* stands for the political or geographical entity, and has been marked as underspecified by backoff.

(4.8) a) LITERAL: La más ambiciosa de aquellas expediciones fue un viaje hasta *América* y a través del Pacífico [...]
(The most ambitious of those expeditions was a journey to *America* and through the Pacific [...])

b) METONYMIC: Los lituanos decidieron elegir al noble alemán Wilhem de Urach como Mindaugas II, pensando que así *Alemania* dejaría más rápida el país obteniendo de esta forma la deseada independencia.
(Lithuanians chose to elect the German as nobleman Wilhem of Urach as Mindaugas II, thinking that in this way *Germany* would leave the country faster, thus obtaining their desired independence.)

c) UNDR-PLURALITY: Es el español que peor se ha adaptado a *Inglaterra*, o que Inglaterra menos ha sabido aprovechar.
(He is the Spaniard that has adapted the worst to *England*, or that England has been less able to take advantage of.)

d) LITERAL: *Alemania* es el país con más vacaciones al tener 42 días por año.
(*Germany* is the country with most holidays having 42 days per year)

We can interpret the low-agreement examples that get tagged as underspecified by backoff in a manner similar to way we interpret the plurality-marked underspecified examples. We have not found any qualitative difference in the examples that are marked as underspecified by plurality or by backoff.

4.10 Summary

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 and *moderate* ($\alpha > 0.41$) for four. Not all the dot types are equally easy to annotate. The main source of variation in agreement, and thus annotation reliability, is the difficulty to identify the senses for each particular dot type. While ENG:ANIMEAT and ENG:LOCORG appear to be the easiest, ENG:ARTINFO and ENG:PROCRES obtain very low α scores.

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 comparable to that of our volunteer annotators. However, to assess the reliability of these sense distributions, we need to look at the agreement values for each dataset, which are provided in Table 4.8.

We assign sense tags using two different methods, VOTE and MACE. VOTE is a rule-based majority voting with a theory-compliant backoff and MACE is an unsupervised EM method that assigns weights to annotators. After evaluating the sense assignments they generate, we set VOTE as a reference SAM.

4.11 Conclusions

We have obtained human-annotated data from turkers for English, and from volunteers for Danish and Spanish. We have compared the sense distribution between turkers and volunters, and between turkers and a single expert voting.

Turkers have a dispreference for the underspecified sense tag, and there are very few (never over 0.01%) of the examples that receive an underspecified sense tag from simple majority.

In the volunter-annotated datasets, however, there is between 8.80% and 10.80% for the total examples for each dataset that receive an underspecified example. If a value over 5% is significant, we can determine that volunteer annotators can identify the underspecified sense.

Comparing the behavior between turkers and volunteers we can determine that the underspecified sense cannot be explicitly captured by turkers using the annotation procedure, and it requires using a sense assignment method (SAM) to estimate which examples are candidates for the underspecified sense. Volunteers, however, agree on the underspecified sense on 10% of the examples. This indicates that human annotators without a bias against the underspecified sense can recognize it, in spite of language differences.

We have compared two SAMs: a voting system (VOTE) with a backoff to the underspecified sense when there is full disagreement between the literal and metonymic senses, and an unsupervised EM method that assigns weights to annotators (MACE). We have chosen the VOTE scheme as SAM to generate the final sense tags for our data because it provides distributions that are more similar to the ones obtained from the expert voting.

Chapter 5

Word sense induction

This chapter is the first of the chapters that describe the NLP experiments we have conducted on the human-annotated dot-type data described in Chapter 4. In this chapter we describe a Word Sense Induction (WSI) experiment on dot type words. WSI relies on a Distributional Semantic Model (DSM) over a large corpus to cluster the contexts of the analyzed words. The centroids of the resulting clusters are the distributional representation of the induced senses.

We carry out this experiment to determine whether there is enough distributional evidence to identify the literal, metonymic and underspecified senses of dot types. In particular, the goal of the experiment is to determine whether there is enough distributional evidence to identify the underspecified sense.

WSI is the token-based homologue to the type-based modelling of regular semantics we cover in Section 3.3, to the extent that type-base modelling determines the classes a word belongs to at the type level, and WSI determines the senses a word belongs to at the token level. For more on WSI, cf. Navigli (2009); Manandhar et al. (2010); Navigli and Vannella (2013b).

We use WSI to determine whether we can infer the literal, metonymic and underspecified senses from corpus in an unsupervised fashion. To determine the validity of the induced senses, we use the data described in Chapter 4 as gold standard at evaluation time. We expect that the higher-agreement data sets provide higher homogeneity (cf. Section 5.3) results because their annotations are more consistent.

This is a class-based approach, and we replace all the words in a dot type for a placeholder, in order to induce senses for a whole dot type at once. Our claim in doing so is that all words belonging to a dot type can manifest the same kind of regular polysemy, and therefore their senses can be modelled together as the common senses of a semantic class or type.

Our system employs the five-word window contexts of each dot-type placeholder to induce the different senses. Once the senses are induced, we assign a sense to each example in the test data, yielding clusters of examples along the induced senses.

This chapter is an expansion of the work documented in Romeo et al. (2013), where the authors only implement WSI for the English data. In this chapter we also use WSI to induce senses for the Danish and Spanish datasets. Other than that, the method followed in this chapter is the same as the one we use in in Romeo et al. (2013).

5.1 Preprocessing

WSI makes use of DSMs to induce word senses. DSMs require large corpora to yield sense vectors we can cluster senses from. For our experiments we used the UkWac corpus for English (Baroni et al., 2009), KorpusDk for Danish (Andersen et al., 2002) and IulaCT for Spanish (Vivaldi, 2009) to fit our WSI models.

After lemmatizing, lowercasing and removing all punctuation from the corpora, we trimmed down the corpora to keep only the sentences that were at least five tokens long. Moreover, since UkWac is very large, we have used a randomized subset instead of the whole corpus for memory reasons. Table 5.1 shows the number of sentences and tokens for each corpus after preprocessing.

Corpus	Sentences	Tokens
Danish	2.7M	44M
English	2.8M	60M
Spanish	1.15M	21M

Table 5.1: Number of sentences and tokens for each corpus for WSI

Grammatical words are important cues for the resolution of metonymies, and we do not want our system to rely solely on content words. We did not remove any stop words from the corpora because we expect prepositions and articles to be important features to distinguish the different senses of dot types.

Regular polysemy is a class-wise phenomenon (cf. Section 2.2), so we expect that all words in a dot type will predicate their literal, metonymic and underspecified senses in a similar manner, i.e. in similar contexts. Thus, our intent is to induce the same senses for all the words of a given semantic class, making our approach class-based.

To group the occurrences of all words of a given dot type, we replaced their occurrences with a placeholder lemma that represents for the whole dot type (*animeatdot, artinfodot, contcontdot, locorgdot, procresdot*). For instance, the lemmatized examples a) and b) with the headwords *paris* and *london* become the sentences in the examples c) and d). Notice how all the instances of *london* become *locorgdot*:

(5.1) a) whilst i be in **paris** in august i decide to visit the catacomb

b) you can get to both **london** station on the **london** underground

c) whilst i be in **locorgdot** in august i decide to visit the catacomb

d) you can get to both **locorgdot** station on the **locorgdot** underground

Using placeholder lemmas has two motivations, namely a theoretical and a technical one. The technical motivation for using placeholders is that they allow us to represent all the words of a dot type under one symbol, which allows class-based WSI. The technical motivation is that, in doing so, we also compensate for sparseness in training and testing data.

Replacing individual lemmas by a placeholder for the overall class yields results similar to those obtained by building prototype distributional vectors for a set of words once the DSM has been calculated (cf. (Turney and Pantel,

2010) for more on prototype vectors of a semantic class). Prototype vectors for a semantic class in a DSM are constructed by adding the vectors for the members of the class; e.g. a prototype vector for ARTIFACT•INFORMATION would be built by summing up the vector for *book*, *novel*, *dictionary*, etc. However, constructing the prototype vector for a class-based induced sense is not trivial because we need to know which vectors to add together from the k induced senses for each member of the semantic class.

To avoid this complication, our take is a preprocessing of the corpus to assure we induce senses directly for the placeholder lemmas. In this way, we avoid having to reconstruct overall class-wise senses from the induced senses for all the individual lemmas. Thus, the placeholders represent the entire dot type from the beginning of the process, as opposed to prototype vectors, which would have to be constructed post hoc.

Using placeholder lemmas to stand for the whole class also provides the added benefit of circumventing some data sparseness, especially for evaluation purposes. For instance, in our data there are some lemmas (e.g. *anchovy, yak, crayfish* in ENG:ANIMEAT) that only appear once in the gold standard. If there is only one example for *yak* in the gold standard, we cannot evaluate the quality of all the induced senses of *yak*. Using placeholder substitution reduces the impact of this sparseness on evaluation by considering each individual lemma as an instance of the dot type the placeholder represents. In other words, we induce senses for *animeatdot* and assign them to occurrences of *yak*, *lamb*, or any other word of the ANIMAL•MEAT dot type.

This replacement method is not exhaustive because we strictly replace the words from the test data by their dot-type placeholder and, for instance, plenty of country and city names are not replaced by *locorgdot*. Appendix A.1 lists the words for each dataset.

5.2 Applying WSI

Our WSI models were built using the Random Indexing Word Sense Induction module in the S-Spaces package for DSMs (Jurgens and Stevens, 2010). Random Indexing (RI) is a fast method to calculate DSM proven to be as reliable as other word-to-word DSMs, like COALS (Rohde et al., 2009). In DSMs, words are represented by numeric vectors calculated from association measures of the analyzed word and their contexts. The similarity between words is measured by the cosine of the vectors of the words being compared.

In a WSI scheme, instead of generating one vector for each word, each word is assigned k vectors, one for each induced sense. These k vectors are obtained by clustering the contexts of each analyzed word into k senses. In our approach, the features used to cluster the contexts into senses are the five words to the left and right of the target word. The output of the system is a DSM where each vector is one of the k-induced senses for the required words. In our case, the words that receive the induced senses are the placeholder dot-type lemmas *animeatdot, artinfodot*, etc.

We use K-means to cluster the contexts into induced senses. Fitting this clustering algorithm requires setting a value of k clusters. We run the WSI system using three different values of k. Ideally the system will work best for $k = 3$, and each of the induced senses will be the literal, metonymic and

underspecified sense respectively.

We want to compare this ideal parameter setting with a coarser and a finer-grained solution, and pick two more values of k, namely 2 and 6. The $k = 2$ solution will work best if each induced sense represents the literal or metonymic sense and there is distributional evidence for the underspecified sense. The $k = 6$ will work best if there is enough distributional evidence to identify more senses that the three senses we have obtained in the annotation task in Chapter 4.

5.3 Evaluation

After fitting a WSI model to obtain the induced senses from the input corpus, we need to evaluate the quality of the induced senses. We do so by assigning the human-annotated data described in Chapter 4 with the induced senses. The S-Spaces package permits the calculation of a vector in a DSM for a new, unobserved example. For each sentence in the test data, we isolated the placeholder to disambiguate, and calculated the representation of the sentence within the corresponding WSI model using the specified 5-word context window.

Once the vector for the sentence was obtained, we assigned the sentence to the induced sense representing the highest cosine similarity out of the k available for each model. This is an unsupervised task, and we refer each to assignment of induced senses to examples in the test data as a *clustering*. Each clustering was evaluated by comparing the received senses from the WSI with the expected senses assigned by VOTE.

Clusterings differ in two parameters: the dataset being used as test data, and the value of k. To measure the quality of the clusterings we use the information-theoretic measures of *homogeneity*, *completeness* and *V-measure* (Rosenberg and Hirschberg, 2007), which have been used as metrics for the SemEval sense-induction shared tasks (Manandhar et al., 2010). These three measures compare the output of the clustering with the expected classes of the test data, and provide a score that can be interpreted in a manner similar to precision, recall and F1, respectively.

Homogeneity determines to which extent each cluster only contains members of a single class, and completeness determines if all members of a given class are assigned to the same cluster. Both the homogeneity and completeness scores are bounded by 0.0 and 1.0, with 1.0 corresponding to the most homogeneous or complete solution respectively.

V-measure is the harmonic mean of homogeneity and completeness. Values close to zero indicate that two label assignments (e.g. the clustering and the gold standard) are largely inconsistent. Much like F1, the V-score indicates the trade-off between homogeneity and completeness. We cannot know a priori whether the sense distribution of the test data and the training data are similar, but we assume they are.

Other WSI tasks (Artiles et al., 2009; Martin-Brualla et al., 2010) make use of two baselines (ONE-IN-ALL and ALL-IN-ONE) to compare to the results of their systems. The ONE-IN-ALL baseline gives each example its own cluster and generates maximum homogeneity and minimum completeness. The ALL-IN-ONE BASELINE places all the examples in the same cluster and provides minimum homogeneity and perfect completeness. These baselines are useful when using

non-parametric clustering, where the systems have to discover an appropriate value of k. Using these baselines is not straightforward for this method.

In our case, we are using a parametric method and the ONE-IN-ALL baseline is not implementable because we are using a fixed value of k. It is a baseline designed to penalize homogeneity and maximize completeness, and, with the same intention, we use a random baseline instead. We do not provide the values of the random baseline (RBL) because they are not informative per se.

In this chapter we consider the difference of performance between two systems to be significant if the proportion between the two performance scores falls out of a 95% confidence interval. The systems are compared in Appendix C.

5.4 Results

The main goal of this experiment is to capture the sense alternation of dot types by computational means, in particular to isolate the distributional behavior of the underspecified sense. To test this we employ a WSI system to induce the senses and subsequently cluster dot-type nominals into three different solutions ($k = 2, 3, 6$).

When evaluating on the data with sense tags obtained using VOTE SAM, most of the clusterings are significantly better than the random baseline, regardless of the value of k. For $k = 2$, ENG:LOCORG is significantly worse than the RBL, and ENG:ANIMEAT is insignificantly worse. For $k = 3$, only ENG:ANIMEAT is worse than RBL.

Table 5.2 presents the results in terms of their homogeneity, completeness and V-measure with a fixed value of $k = 3$ for each dataset.

Dataset	HOM	COM	V-ME
ENG:ANIMEAT	0.0055	0.0033	0.0041
ENG:ARTINFO	0.0213	0.0190	0.0200
ENG:CONTCONT	0.0290	0.0197	0.0235
ENG:LOCORG	0.1067	0.0785	0.0905
ENG:PROCRES	0.0049	0.0042	0.0046
DA:CONTCONT	0.0080	0.0069	0.0074
DA:LOCORG	0.0386	0.0490	0.0431
SPA:CONTCONT	0.0013	0.0012	0.0013
SPA:LOCORG	0.0105	0.0120	0.0111

Table 5.2: VOTE: Results for the $k = 3$ clustering for VOTE in terms of homogeneity (HOM), completeness (COM) and V-measure (V-ME)

In Figures 5.1 and 5.2 we depict the variation in V-measure for each VOTE dataset as k increases. The V-measure values are low, with ENG:LOCORG showing the highest performance at 0.11. ENG:LOCORG is the best English dataset for $k = 3$, and the only that worsens significantly for $k = 6$.

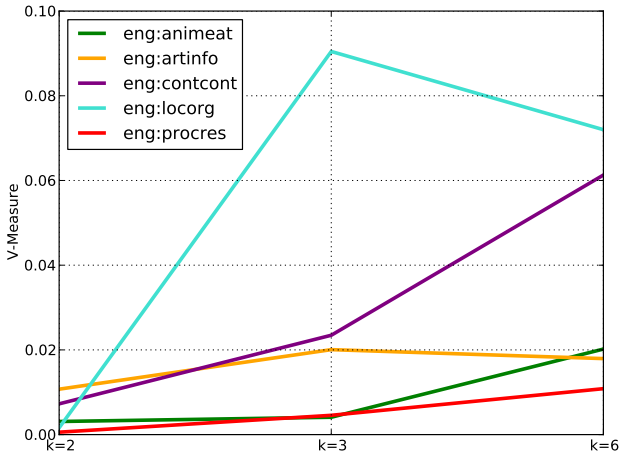

Figure 5.1: V-measure for the English datasets for $k = 2, 3, 6$

The improvement of performance in terms of V-measure as k increases is in general non-monotonic. This improvement is not correlated with the agreement of the datasets, but rather a consequence of the higher homogeneity obtained when clustering with a higher value of k.

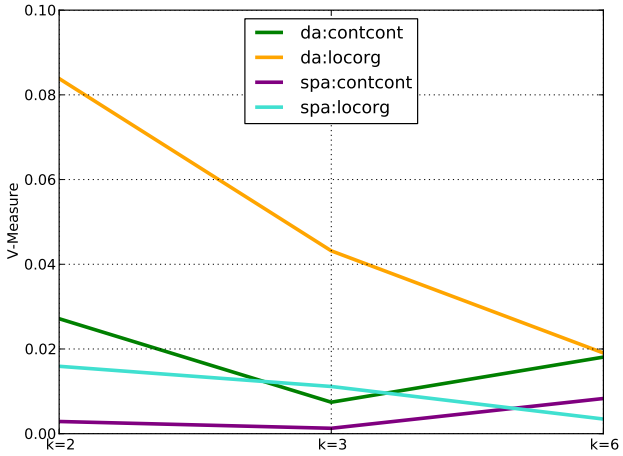

Figure 5.2: V-measure for the Danish and Spanish datasets for $k = 2, 3, 6$

The datasets for the CONTAINER•CONTENT dot type for Danish and Spanish improve as k increases. This tendency towards improvement is congruent with most of the English datasets. However, the performance for the LOCATION•ORGANIZATION datasets worsens with the increases in k. Again, the improvement is not correlated with agreement.

5.4.1 Feature analysis

The performance scores for the clusterings are low, but improve over RBL. This indicates that this WSI method captures some distributional information useful in identifying some of the sense alternations in the datasets.

The underspecified examples tend to be evenly spread across the clusters. Distributional evidence does not spawn a mostly underspecified cluster. Table 5.3 shows the contingency matrices for ENG:ANIMEAT and ENG:LOCORG for $k = 3$. In ENG:ANIMEAT, the distribution of literal, metonymic and underspecified senses is similar for each of the three clusters. This indicates that the induced senses for the placeholder *animeatdot* fail at capturing the three-fold sense difference we expect.

The confusion matrix is different for ENG:LOCORG. The clusters labelled 0 and 2 have similar proportions of literal and metonymic examples, but cluster $c = 1$ is mostly (87%) made up of literal examples. This indicates that there is indeed some linguistic evidence the WSI system learns to discriminate dot-type senses with.

	ENG:ANIMEAT			ENG:LOCORG		
	L	M	U	L	M	U
c=0	110	51	3	62	69	8
c=1	127	43	1	151	17	5
c=2	121	41	3	94	85	9

Table 5.3: $k = 3$ solutions for ENG:ANIMEAT and ENG:LOCORG dot types

Table 5.4 lists the ranked ten most frequent words for each cluster for ENG:ANIMEAT and ENG:LOCORG for $k = 3$. Most of the words in each cluster are prepositions and articles, which we did not remove form the corpora before running WSI. Notice how the rankings differ, in spite of the low lemma-wise variety,.

In ENG:ANIMEAT, *fish* is one of the most frequent words for $c = 0$, and the placeholder lemma for the dot type, *animeatdot* appears in two of the clusters. This presence of the placeholder lemma is an indication of enumerations like 'sardine, eel and salmon", which would be replaced by "animeatdot, animeatdot and animeatdot". The word *fish* was not replaced by its placeholder as it does not appear in the gold standard data but is one of the few content words in the top 10 words for each cluster.

For ENG:ANIMEAT the articles *the* and *a* were the most frequent words for each cluster, whereas for ENG:LOCORG, the most frequent words that contribute to each cluster are prepositions. In particular, the preposition *in* is the key feature that makes the cluster $c = 1$ mostly literal, because it helps cluster all the examples that contain structures like "*in* Paris" that are literal more often than not.

In c=2, the most important preposition is *to*, which indicates directionality ("I am moving *to* London") and seems a priori a good feature for the literal LOCATION sense. However, this very frequent feature appears in the same slot as *in*, and gets pooled together with other features that make c=1 and c=2 different during the sense-induction process. This suggests that this method generates

vectors for the induced senses that are largely orthogonal to our expected senses.

	ENG:ANIMEAT	ENG:LOCORG
c=0	*and*, *animeatdot*, *of*, *a*, *for*,*with*, *the*, *fish*, *in*, *to*	*the*, *of*, *and*, *to*, *a*, *in*, *that*, *time*, *it*, *for*
c=1	*the*, *of*, *and*, *in*, *a*, *to*, *is*, *that*, *animeatdot*, *with*	*in*, *the*, *and*, *to*, *a*, *of*, *that*, *is*, *it*, *for*
c=2	*a*, *of*, *to*, *in*, *that*, *or*, *is*, *with*, *for*, *from*	*to*, *and*, *from*, *a*, *locorgdot*, *the*, *that*, *with*, *for*, *is*

Table 5.4: Top 10 most frequent context words per c used in $k = 3$ for ENG:ANIMEAT and ENG:LOCORG datasets

The importance of the article as a feature reflects that the mass/count distinction is a key component in the sense alternation of some instances of regular polysemy (such as ENG:ANIMEAT). By lemmatizing, we are discarding potentially relevant morphological information that can help disambiguate the mass/count readings of nouns. For English and Spanish, plural is an important cue to determine whether a noun is count or mass, which also corresponds with the senses of some dot types being either count or mass (cf. Section 2.6). For Danish this is even more relevant because definiteness is also a nominal inflexion besides plural, and also plays a role in determining count or mass readings.

Our system makes no difference between words at the left and the right of the headword. This results in the clustering pooling together nouns headed by the proposition *of* with nouns with a modifier headed by *of*.

The distributional evidence available to the $k = 3$ solution is again not strong enough to motivate an individual cluster for each sense. However, under the assumption that more fine-grained patterns may indicate underspecified readings; we attempted a $k = 6$ solution to differentiate between senses with a larger k.

5.4.2 Inducing a different number of senses

The goal of the $k = 3$ solution is to obtain clusterings that map to the three proposed senses for a dot type (literal, metonymic and underspecified). We have trained the system for $k = 2$ and $k = 6$ to observe the differences in clustering behavior and how they relate to the annotated senses of the datasets.

Inducing two senses

The $k = 2$ solution attempts to mirror a literal vs. metonymic partition between the senses of each dot type.

In this scenario, the underspecified examples are distributed evenly across the two clusters, and do not lean more towards the literal or the metonymic. We observed, for instance, that the underspecified senses of ENG:ARTINFO occurred often with the headword preceded by *of*, and were clustered alongside the metonymic examples because the preposition *of* is a feature for the metonymic sense. Furthermore, the underspecified examples whose headword is the objects

of a verb like *keep* or *see* where clustered with the literal examples. Thus, underspecified examples are subject to the pull of the literal and metonymic when senses are clustered.

The spread of underspecified examples in two clusters suggest that we can try to find distributional evidence for more senses and expect them to cluster in a more homogeneous fashion.

Inducing six senses

In addition to comparing the performance of clustering for $k = 3$ with a clustering with only two clusters, we also trained and tested WSI models for $k = 6$. The aim of the $k = 6$ solution is to identify fine-grained distinctions between the distributional behavior of the analyzed dot types. In Figures 5.1 and 5.2 we see that most datasets show a higher V-measure for $k = 6$ than their $k = 2$ and $k = 3$ counterparts, but this improvement a consequence of the higher homogeneity expected from an increased k-value.

On one hand, the less homogeneous clusters in $k = 3$ are prone to be split into smaller, more homogeneous clusters in $k = 6$. On the other hand, the more homogeneous sub-clusters in $k = 3$ are preserved in $k = 6$.

(5.2)
1. herb,and,medicine,in,**porcelain**,..., to,nepalese,trader,offering,exotic
2. and,18,th,century,chinese,**porcelain**,...,along,wall,hung,with,brussels
3. mammary,cell,grown,on,**plastic**,...,were,unable,to,differentiate,by
4. 500,ml,**plastic**,...,in,styrofoam,cooler,

Example 5.2 shows the 5-word windows for four examples of the ENG:CONTCONT dataset. In the $k = 3$ solution, these four examples fall into the same cluster. However, in the $k = 6$ the words marked in bold make them different enough for the WSI system to place them in two different clusters, one for *plastic* and one for *porcelain*.

The $k = 6$ solution is thus a further refinement of the $k = 3$ into more fine-grained distributional behavior, but still largely orthogonal to our expectations of sense selection for dot types.

5.5 Conclusions

In this chapter, our objective was to use WSI to capture the sense alternation of dot types. The low V-measure of the induced-sense clustering solutions demonstrates that our method was not able to isolate the literal, metonymic and underspecified senses. Our results do not identify an absolute distinction between the senses of a dot type.

In Section 5.4.1 show we have also been able to identify the contribution to sense selection for some high-frequency words, but our power to draw conclusions is limited by the performance of the system. If we achieved a WSI system that recognized the literal and metonymic senses with good performance, but is unable to identify the underspecified sense, we could argue against the underspecified sense as a distributional phenomenon.

The system could be improved by using more refined features. Using a 5-word window of lemmas around the headword seems to generate a limited feature space that does not give account for regular polysemy.

A way of including lemma, form, and syntactic information would be using dependency-parsed input for building of DSMs. By doing this, the WSI could infer senses from more structured and informed input that would incorporate word order, inflectional morphology and lemmatized information. The S-Spaces package allows using parsed corpora as input for DSM generation.

In addition to using more linguistically informed input, other clustering algorithms instead of K-means could be used to induce senses. The system could be extended by using other clustering algorithms like that are non-parametric and do not require a value of k, but rather estimate an appropriate number of senses to fit the data.

The test data we use also biases our result in two ways: by constraining the choice of words we train from, and by influencing our ability to draw conclusions from data with contested reliability.

If we only replace the lemmas from the test data by the placeholder, we are making the system less robust as a class-based WSI. We should expand these lists with more lemmas—e.g. using a thesaurus—, so the distribution of the semantic class can be less biased by the choice of lemmas.

Finally, our evaluation metrics depend not only on an accurate induction of the senses in context, but also on the reliability of the test set. However, we have seen that agreement of the test data does not necessarily correlate with performance of the WSI. Nevertheless, this method is a possible baseline for WSI for regular polysemy, and, to the best of our knowledge, the first attempt to use class-based WSI.

With regards to the objective of identifying the underspecified sense, we have been unable to infer an underspecified sense from the distributional behavior of dot types. In fact, we have not been able to isolate the literal and metonymic senses either. Thus, we affirm that, using this setup for WSI, we have not found distributional evidence to postulate an underspecified sense.

Chapter 6

Features

We concluded Chapter 5 with the remark that a five-word context window generates a very limited feature space for the induction of the senses of dot-type nominals. In this chapter we describe the features we have extracted to characterize the linguistic information of our sense-annotated data. These features attempt to capture more refined linguistic information than the shallow features yielded by a five-word window around the headword.

Boleda et al. (2012a) mention that feature representation typically used by machine learning algorithms provides the empirical handle to the linguistic properties of words. They explicit two preconditions for a lexical modelling task that describes regular polysemy: a) a classification that establishes the number and characteristics of the target semantic classes b) a stable relation between observable features and each semantic class. We cover a) in Section 2.6 and expect b) to be the features that we employ as explanatory variables in the supervised experiments in Chapters 7,8 and 9.

The choice of features is a crucial part in the design of any Natural Language Processing (NLP) experiment, and can be influenced by the availability of linguistics resources for the language or task at hand, the technical constraints of the system and the specific linguistic phenomena that need to be modelled. The criteria for choosing the following features have been generality, theory compliance and comparability with the state of the art.

For the sake of **generality**, there has to be a common feature space for all the semantic classes within a language. Such common space will enable the comparison between semantic classes during the evaluation and the realization of experiments using all the datasets simultaneously. Using a general feature space will increase the overall amount of features, as some will be sparse when training for a certain dataset and more populated for another.

We rule out the possibility of tailoring a specific feature set to a certain semantic class. Our work in Martínez Alonso et al. (2011) uses Sketch Engine (Kilgarriff et al., 2004) to select the relevant lexical elements that are associated to a semantic class in order to build a bag of words. Using this approach would the size of the bag of words, but there would be one bag of words for each dataset.

In terms of **theory compliance**, one of the goals of this dissertation is the empirical evaluation of the theory on dot types. The objective requires the data to be characterized morphosyntactically, as the work on regular polysemy, both

in and outside the GL framework, emphasizes the relevance of morphological factors like *definite vs. indefinite* or *plural vs. singular*, as well as syntactic factors like copredication or argument saturation (Pustejovsky, 1995; Markert and Nissim, 2009; Grimshaw, 1991) for establishing the different metonymic readings.

State-of-the art experiments need to be reproduced for the sake of **comparability**. Some of the features in the following sections have been replicated from relevant approaches to metonymy resolution (cf. Section 3.4). Moreover, replicating features from these experiments allows us to implement the SemEval2007 baselines we describe in Appendix D.

Our features need to comply with the **class-based** approach we have taken in the annotation task—by using one common sense inventory for each dot type—and in the WSI experiments—by using a placeholder lemma for each dot type. We want the system to generalize over particular lemmas, and this requires we exclude the identity of the headword from the learning algorithm when training in order to reduce the bias of the sense distribution for each particular lemma of a dot type. Instead, we focus on characterizing the grammatical and semantic properties of the sentence the headword appears in.

Some of the features depend on Brown clustering or topic models, which are generated in an unsupervised manner. Using unsupervised machine learning to aid supervised learning makes our approach **semisupervised**.

By *grammatical* features we refer to the inflectional characteristics of the headword, plus the syntactic relation of the headword with regards to the other words in the sentence. By *semantic* features we refer to other traits that are not immediately grammatical like the presence of certain semantic types (verb of motion, names of plants) in the context of the headword. We can obtain these features using clustering measures or querying a LKB.

As commented in Section 3.6, we use no selection restriction modelling because metonymy is too conventional for this to make any sense if the selectional restrictions are obtained from corpus—and no selectional restriction violation would be found in very common metonymies like “eating chicken”.

6.1 Grammatical features

This section describes the features that characterize a example in terms of the morphological and syntactic properties of the headword and its related words. For each sentence with a nominal headword h belonging to a dot type, we generate the features detailed in the following sections. Most features are binary unless otherwise noted.

We obtain the morphological features from the part-of-speech tags of h. For English and Danish we use the POS tags the corpora (ANC and KorpusDK) are distributed with. For Spanish, we tokenized and tagged the IulaCT corpus using Freeling (Padró and Stanilovsky, 2012) to make sure the corpus had the same format as the treebank.

For the other grammatical features, we use dependency parsing. Dependency parsing generates a tree structure where each word points to its syntactic head and the edges between words are labelled with syntactic information like *subject* or *subordinating conjunction*.

We have using Bohnet's graph-based parser (Marimon et al., 2012) to generate the dependency trees for our datasets. We have trained the parser on the Danish Dependency Treebank (Kromann, 2003), the IULA treebank (Marimon et al., 2012) for Spanish, and the dependency conversion of the Penn Treebank in (Nugues and Heiki-Jaan, 2007) for English. In Table 6.1 we show the Labelled and Unlabelled Attachment Score (LAS and ULA) for each language on the test data for each treebank.

Language	LAS	ULA
Danish	86.82	91.23
English	91.13	92.35
Spanish	94.68	97.21

Table 6.1: Performance for dependency parsers

The parsing results are high, in particular for the Spanish treebank, which is large and made up of mostly technical text. However, we know that the dependency conventions of treebanks—what is a head of what—have an impact on their learnability by parsers, and on the usefulness of the features extracted from such predicted dependency trees (Schwartz et al., 2012; Elming et al., 2013).

In particular, for the Danish treebank, articles are the heads of nouns. Thus, for a phrase like "på det danske hold" ("in the Danish team"), the dependency tree is the one shown in Figure 6.1. The arrows represent the dependency relation between words, where the outgoing node is the head and the incoming node is the dependent. The labels on the arrows represent the syntactic role assigned to the dependent[1]

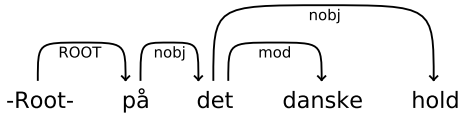

Figure 6.1: Original dependency structure with article as head

However, this structure is not convenient if we want to identify the verbs or preposition that the nominal headword h depends on, because they are placed as grandparents and not as parent nodes. For this reason, we have rewritten the Danish dependencies to make the article (*det*) the dependent of the noun (*hold*). In this way, the noun that becomes the head is now the dependent of the previous head of the article (the preposition *på*), and the adjective (*danske*) becomes the dependent of the noun. The resulting structure, shown in Figure 6.2, is the same structure provided by the English and Spanish treebanks.

[1]The trees have been plotted using *What's Wrong With My NLP?* from http://code.google.com/p/whatswrong/

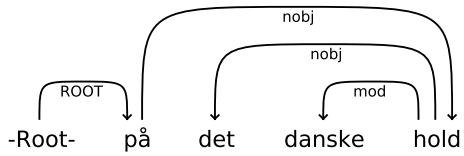

Figure 6.2: Modified dependency structure with noun as head

6.1.1 Inflectional morphology

These features list the inflectional traits of h. We obtain these features from the part-of-speech tags of the h. For English and Spanish, there is only one feature to discriminate between plural and singular. For Danish, in additional to plural, one feature reflects the definite/indefinite alternation, and another feature the genitive case marking, if present (cf. Section 2.6.6).

We group the inflection-morphology related features in a feature group `m`. Table 6.7 shows the different feature groups for each language.

6.1.2 Syntactic roles

We use the dependency labels from the parsed tree for each sentence to determine the syntactic role of h, its head and dependents. The syntactic role of h generates two baselines for metonymy resolution (cf. Appendix D). We refer to the feature group for the possible values of the dependency label of h as `pls`, to the features for the dependency label for the head as `plh`, and to the features for the dependency label of the dependents of h as `plc`.

Each treebank has a different amount of dependency labels but all provide key syntactic traits such as whether a word is subject or object. The list of total labels shown in table 6.2 is obtained from possible labels in the English training data. Each of the feature groups use a subset of this list, limited to the labels found for the set of possible labels h (`pls`), the head of h (`plh`), or its dependents (`plc`).

ADV	AMOD	APPO	CONJ	COORD	DEP
DIR	GAP-SUBJ	HMOD	HYPH	IM	LGS
LOC	LOC-PRD	MNR	NAME	NMOD	OBJ
OPRD	P	PMOD	POSTHON	PRD	PRN
PRP	PRT	ROOT	SBJ	SUB	SUFFIX
TITLE	TMP	VC			

Table 6.2: List of dependency labels for English

6.1.3 Head and dependents

Syntactic features like grammatical relations have proven useful for general word sense disambiguation (Martínez et al., 2002) and metonymy resolution (cf. Section 3.4). In Rumshisky et al. (2007), verbs and adjectives that are syntactically related to a dot-type noun are defined as *selectors* for the different senses of a dot type.

Even though we are not using a coercion or selectional-preference approach to our feature modelling, we acknowledge the importance of noting the syntactically related words to h. We include two sets of feature groups, one depending on the lemmas and the other depending on the POS-tags of the syntactically related words.

We include the feature group `phw` to give account for the lemmas of different syntactic heads of h, and `pcw` for the lemmas of the dependents of h. The two groups that depend on POS-tags are `ppc` and `pph`, which list the POS-tags of the head and dependents of h. In Section 6.2.2 we describe another feature group that lists clustered head and dependent lemmas.

6.1.4 Syntactic counts

The feature group `pq` is made out of three features that represent counts instead of binary values. These features aim at framing h in a more general, sentence-wise way than the other syntactic features:

1. The number of dependents of h
2. The number of siblings of h, i.e. how many other dependents the head of h has.
3. The distance from h to the root node of the dependency tree, normalized over the total number of words in the sentence.

6.2 Semantic features

6.2.1 Bag of words

The simplest way of incorporating lexical information, a bag of words (`bow`) is a set of features in which each individual feature corresponds to a word in the vocabulary. We use a bag of words with all content words from the training corpora.

Even though in 5.4.1 we keep prepositions and determiners, we do so because we are using a restricted word window, and we expect that most stop words that appear in that window will be related to the headword. Keeping a bag of words over an arbitrarily long sentence provides no guarantee that determiners and prepositions will be indeed related to the headword, and we discard them.

Note how the `bow` feature group is much is much larger for English than for the other languages. This is a result of using five datasets for English instead of two. The size of the bag of words grows almost linearly for each new dataset.

6.2.2 Brown clusters

The Brown algorithm (Brown et al., 1992) takes as input a corpus and produces a hierarchical clustering of the vocabulary of the corpus. The algorithm initializes by setting each word in the vocabulary in its own individual cluster, thereby starting with as many clusters as types. Each iteration merges together the pair of clusters which minimizes the decrease of likelihood of the corpus if joined. This likelihood is measured using a bigram model.

The result of the merge operations can be represented as a binary tree. Any given word in the vocabulary is thus identified by its path from the root node, as seen in Figure 6.3.

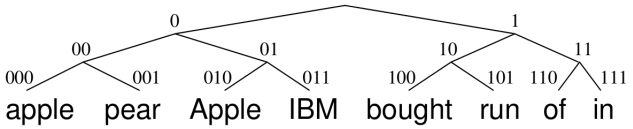

Figure 6.3: Example of Brown clustering from Koo et al. (2008)

We use the implementation by Liang (2005) to generate Brown clusters from our corpora for all three languages using the default settings of 1000 clusters[2].

This clustering method generates clusters that are semantically coherent but contain interferences. Table 6.3 shows an example of words clustered in three different but related clusters. Only the top-frequency words for each cluster are shown. The first cluster shows linguistic units like *noun* or *verb*, which are distributionally similar to each other, and also *phrase* and *word*. The second cluster has related words like grammar, dialect and language, and other less expected words like *landscape* or *universe*. Likewise, the third cluster contains words related to documents of some kind (*album, diary, book, poem*) and the word *millennium*, which the brown algorithm has clustered together by virtue of its contextual similarity to the other words (possibly being introduced by the possessive *our*).

[2]https://github.com/percyliang/brown-cluster

Cluster	Word
00001011100	...
00001011100	noun
00001011100	verb
00001011100	phrase
00001011100	word
00001011101	grammar
00001011101	dialect
00001011101	currency
00001011101	landscape
00001011101	universe
00001011101	usage
00001011101	text
00001011101	truth
00001011101	language
00001011111	album
00001011111	diary
00001011111	poem
00001011111	book
00001011111	millennium

Table 6.3: Example clusters from the ANC

All cluster identifiers in this example are eleven bits long. If we drop the last two tailing bits and keep only the common **000010111**, all the words fall within the same cluster. Different prefix lengths allow different clustering granularities. Less grainy clusters will be less numerous and contain more words, and thus more noise: in this example, cluster **000010111** contains *book* and *language*, but also *millennium*, *landscape* and *truth*.

Clustering bag of words

An alternative to using a bag of words as a feature set is using Brown clustering to reduce the amount of types in the vocabulary of the corpus. This has the additional benefit of capturing all the words in the corpus, as they are all assigned to a Brown cluster, something that does not happen if only the top N words of a corpus are used to build the bag of words.

We generate a Brown-clustered bag of words (the feature group `bb`) for each language from the lemmatized content words from the training data. The clustering algorithm is set to the default configuration of 1000 clusters. Using Brown clustering collapses the bag of words into a smaller feature group. For instance, the words *landscape*, *universe* and *dialect* would be assigned to the feature `bb:00001011101` instead of each having an independent feature.

Clustering heads and dependents

Following the work in Nissim and Markert (2003), we reduce the amount of head or dependent features by clustering the lexical heads and dependents. Instead of using a thesaurus, we use Brown clustering from this section to replace each

lemma by the identifier of its clusters, thus reducing the cardinality of the feature groups. This generates the feature groups `bh`, for the clustered syntactic heads of h, and `bc` for the dependents of h.

6.2.3 Topic models

Topic models characterize any document as a probabilistic mixture of different topics. Topic models are generated by fitting a corpus made up of documents over an arbitrary number of topics, where each word as a probability of being generated by one of the topics that compose the document. After running the MALLET topic-model package McCallum (2002) over 100 topics on the ANC, we obtained topics like the ones shown in Table 6.4.

Index	Words
4	gene protein sequence genome domain group tree bacteria ...
5	black american white african action race class affirmative ...
6	make head hand eye back water man dog eat leave foot ...
7	diet food weight animal fish rat **study** birth mouse ...
8	**study** risk estimate analysis effect health benefit factor...
9	time book write magazine editor york story read yorker call ...

Table 6.4: Example topics calculated from the ANC

The different topics are made up of word lists that become thematically grouped according to their co-appearance in the training documents. Topics 4 and 9, for instance, show words that relate to the themes of biology and the publishing world respectively. The other topics are also easy to interpret as encompassing a family of words that relates to a particular subject, often called a *domain*. Some words belong to more than one topic. The word study, for instance, appears in topics 7 and 8.

The role of domain in metonymy resolution has been argued for by (Croft, 1993). Li et al. (2010) use topic models as features for WSD. We use topic models to include domain-related information our feature scheme. The feature group `t` contains the list of 100 probability values of a sentence belonging to each of the 100 topic models we calculated with an Latent Dirichlet Allocation sentence-wise document clustering using MALLET.

6.2.4 Ontological types

In order to incorporate the semantic information available in LKBs, we use the wordnets for Danish, English and Spanish to characterize the semantic types of the context words of h. Each of the binary features in the `wn` group informs on whether there is a content word in the context of h that belongs to a wordnet unique beginner (or ontological type, cf. Table 6.5). The semantic class of each word is obtained from the first synset for that lemma.

ADJ.ALL	ADJ.PERT	ADJ.PPL	ADV.ALL	N.TOPS	N.ACT
N.ANIMAL	N.ARTIFACT	N.ATTRIBUTE	N.BODY	N.COGNITION	N.COMMUNICATION
N.EVENT	N.FEELING	N.FOOD	N.GROUP	N.LOCATION	N.MOTIVE
N.OBJECT	N.PERSON	N.PHENOMENON	N.PLANT	N.POSSESSION	N.PROCESS
N.QUANTITY	N.RELATION	N.SHAPE	N.STATE	N.SUBSTANCE	N.TIME
V.BODY	V.CHANGE	V.COGNITION	V.COMMUNICATION	V.COMPETITION	V.CONSUMPTION
V.CONTACT	V.CREATION	V.EMOTION	V.MOTION	V.PERCEPTION	V.POSSESSION
V.SOCIAL	V.STATIVE	V.WEATHER			

Table 6.5: List of ontological types in Princeton WordNet

We have used the MCR30 (Gonzalez-Agirre et al., 2012) distribution for the English and the Spanish wordnets, which share their list of ontological types. For Danish we have used DanNet (Pedersen et al., 2006), which provides a slightly different set of ontological types (cf. Table 6.6). The ontological types in DanNet no dot explicitly depend on the part of speech, but rather on the order (first, second or third) of the entity they stand for (cf. Section 2.3.2). In the English and Spanish wordnets, there is for instance a cognition type for verbs and another for nouns (v.cognition and n.cognition, respectively).

1STORDERENTITY	2NDORDERENTITY	3RDORDERENTITY	AGENTIVE	ANIMAL	ARTIFACT
ARTWORK	BODYPART	BOUNDEDEVENT	BUILDING	CAUSE	COLOUR
COMESTIBLE	COMMUNICATION	CONDITION	CONTAINER	CREATURE	DOMAIN
DYNAMIC	EXISTENCE	EXPERIENCE	FORM	FURNITURE	GARMENT
GEOPOL.	GROUP	HUMAN	IMAGEREP.	INSTITUTION	INSTRUMENT
LANGUAGEREP.	LIQUID	LIVING	LOCATION	MANNER	MENTAL
MONEYREP.	NATURAL	OBJECT	OCCUPATION	PART	PHENOMENAL
PHYSICAL	PLACE	PLANT	POSSESSION	PROPERTY	PURPOSE
QUANTITY	RELATION	SOCIAL	STATIC	STIMULATING	SUBSTANCE
TIME	UNBOUNDEDEVENT	UNDERSPECIFIED	VEHICLE		

Table 6.6: List of ontological types in DanNet

The feature group `wn` determines whether the sentence h is part of contains words of the ontological types shown in Tables 6.5 or 6.6.

6.3 Feature summary

Table 6.7 shows, for each language, the size of each feature group.

Feat. group	Danish	English	Spanish	Description
bow	4,730	10,332	5,364	lemmatized bag of words
bb	981	976	969	Brown-clustered bag of words
bc	160	473	153	Brown-clustered syntactic dependents
bh	206	416	163	Brown-clustered syntactic heads
m	3	1	1	inflectional morphology og h
pwc	199	679	238	lemmatized syntactic dependents
pwh	280	878	190	lemmatized syntactic heads
plc	17	18	13	syntactic role of dependents
plh	17	18	13	syntactic role of head
pls	13	14	7	syntactic role of h
ppc	14	29	10	POS tag of dependents
pph	9	18	54	POS tag of head
pq	3	3	3	# of children, # of siblings, distance to root
t	100	100	100	topic models
wn	39	45	45	ontological types
ALL	2,041	3,668	1,959	
ALB	6,771	14,000	7,323	

Table 6.7: Size of each feature group for each language

The last two rows show the total size for all the features that depend on an external resource like parsing, a clustering or a LKB (ALL), and for all the features including the bag of words feature group (ALB). ALL and ALB are two of the feature sets listed in Table 6.8.

There are fifteen different feature groups from the grammatical and semantic features. To evaluate how the supervised systems perform on different features, we have pooled combinations of feature groups together into *feature sets*, which we list in Table 6.8. We compare the performance of the different feature sets for the supervised experiments in the following chapters.

	ALB	ALL	BOW	BRW	PBR	PBW	PLB	PLE	PPW	SEM	TOP	WNT
bow	•		•				•					
bb	•	•		•					•	•		
bc	•	•			•	•						
bh	•	•			•	•						
m	•	•			•	•	•	•	•			
pwc	•	•						•				
pwh	•	•						•				
plc	•	•			•	•	•	•	•			
plh	•	•			•	•	•	•	•			
pls	•	•			•	•	•	•	•			
ppc	•	•					•		•			
pph	•	•					•		•			
pq	•	•			•	•	•	•	•			
t	•	•				•			•	•	•	
wn	•	•				•			•	•		•

Table 6.8: Feature sets
(columns) and the feature groups they are made up from (rows).

Chapter 7

Word sense disambiguation

The most immediate NLP experiment to execute when there are linguistic features and sense annotations available is to predict the annotated senses using the linguistic features as explanatory variables. In this chapter we use the features from Chapter 6 to attempt a mapping between the linguistic traits they characterize and the assigned senses of the dot-type datasets by using supervised learning.

In Chapter 5, we have determined that an unsupervised method like WSI with a five-word context window as feature space is not sufficient to capture the sense alternations of the dot types we describe in Chapter 2 and annotate in Chapter 4. In Chapter 6 we define the features we use to replace the overly simplistic five-window context of the WSI experiment. These features incorporate grammatical features that are derived from POS tags or dependency parsing, and semantic features that use word clusterings or LKBs.

We carry out an experiment using supervised machine learning to automatically reproduce the judgments of the annotators on the dot-type datasets from Chapter 4. In particular, our goal is to measure the identifiability of the underspecified sense.

Using supervised learning to assign an item with a category from a discrete set is called *classification*. A classification task that assigns word senses as target variables is an instance of *word-sense disambiguation* (WSD). The goal of applying WSD to our sense-annotated data is to establish the learnability of the literal-metonymic-underspecified sense distinction for each dataset with regards to the linguistic characterization the extracted features provide.

However, some of the features we use depend on Brown clustering or topic models (cf. Chapter 6), which are generated in an unsupervised manner. Using unsupervised machine learning to aid supervised learning makes our approach semisupervised. Besides incorporating unsupervised results in the feature representation, other semisupervised strategies like self-training or co-training can be addressed. However we limit the scope of the semisupervision in our experiments to feature representation and make no use of self-training or co-training, because, in spite of the potential benefits, these techniques might lead to overfitting, and require an additional effort in parametrization. For more on the difficulties of dealing with WSD and self-training, cf. Abney (2010); Mihalcea (2004).

Much like in Chapter 5, our approach is class-based, because regular pol-

ysemy is a class-wise phenomenon and we expect, for instance, the linguistic features of the words *chicken* and *salmon* to be useful in predicting the senses of the words *lamb* and *tuna*, which also belong to the ANIMAL•MEAT dot type.

This chapter is an expansion on two previous class-based WSD experiments that use part of our annotated data, namely Martínez Alonso et al. (2011) and Martínez Alonso et al. (2012).

In Martínez Alonso et al. (2011), we aimed at classifying the literal, metonymic and underspecified senses of the ENG:LOCORG dataset. In an attempt to palliate the effects of the sense distribution being skewed towards the literal (LOCATION) sense, we generated additional training data for the metonymic class by using examples of nouns that were organizations. The F-score for the literal sense was 0.81, but the performance for the metonymic and underspecified senses was much lower, at 0.55 and 0.51 respectively.

In Martínez Alonso et al. (2012), we used an ensemble classifier to try to improve the recognition of the underspecified sense for the ENG:LOCORG and the ENG:CONTCONT datasets. However, the resulting system had an even stronger bias for the most frequent sense than any of the individual classifiers separately.

This chapter expands on these two experiments and provides several methodological changes. First of all, in both experiments we used expert annotations because there were no turker or volunteer annotations at the time. Instead, in the following experiments we use the sense assignments provided by VOTE—which are aggregated from turker or volunteer annotations—to implement our WSD systems.

The main reason to disprefer the expert annotations with regards to a WSD system is that it is more realistic and reproducible to use crowdsourced annotations. Moreover, given the goal of the task—to measure the identifiability of the underspecified sense—, we consider that annotations obtained from more than annotator are more adequate to assess the relevance of the underspecified sense. We give the reasons to prefer the non-expert datasets for our task in Chapter 4.

In addition to having more annotations available, we also have seven datasets more, for a total of nine. The experiments described in this chapter make use of the five English, two Danish and two Spanish datasets, thus implementing WSD for all nine datasets.

Lastly, the features used for WSD in Martínez Alonso et al. (2011) and Martínez Alonso et al. (2012) were shallow features that did not use parsing to obtain word relations, or any kind of clustering to obtain semantic features. For the WSD experiments in this chapter we use the full repertoire of grammatical and semantic features described in Chapter 6.

In Section 7.1 we describe the parameters that our WSD system has to give account for. In Section 7.2 we describe the variation in accuracy baseline for each dataset. In Section 7.3 we describe the evaluation of the WSD system. In Section 7.4 we describe a classifier ensemble aimed at improving the identification of the underspecified sense. Finally, in Section 7.5 we elaborate on what performance of the WSD system indicates about the need to include the underspecified sense in token-wise sense inventories for dot types.

7.1 Method

Before determining whether we can identify the underspecified sense for the annotated datasets, we need to develop a reliable WSD system. In this section we provide an overview of the technical aspects that our WSD has to consider. The system has several parameters to give account for:

1. There are **nine datasets**, five for English and two for Danish and Spanish each (cf. Chapter 4). Each dataset has its own particularities in terms of baselines for WSD, relevant features, distribution of senses, reliability of annotation and expected predictability of senses.

2. Each dataset has **two annotation variants**, one for each SAM. We have determined in Chapter 4 that we will use VOTE. The system is also evaluated on MACE in Appendix E.2 for the sake of reference, but we conduct our study using the VOTE sense assignments.

3. From all the grammatical and semantic features from Chapter 6, we have established **fourteen different feature sets** (cf. Table 6.8) for training and testing of our classifiers. We need to determine the feature set that fares best overall.

4. In order to evaluate the performance of WSD, we need to use a learning algorithm to learn a mapping from the features to the annotated senses. When using datasets of relatively small size like ours, it is desirable to use algorithms with few parameters to tune for. For this reason we have abstained from using using parametrization-heavy learning algorithms like Support Vector Machines, and instead used the default settings of and used **three classifications algorithms** from the SkLearn (Pedregosa et al., 2011) implementation: Naive Bayes, logistic regression and decision trees.

In Section 7.3 we cover the effect of each parameter in the WSD task. In Appendix E we provide the tables that list the accuracy for each dataset, feature set, SAM and learning algorithm.

7.2 Baseline

There are three possible baselines determined by Markert and Nissim (2009) (cf. 3.4), but we only use the most-frequent sense baseline or MFS. In Appendix D we provide a comparison of the baselines and determine that they are in general not harder baselines for this task than MFS.

DATASET	Acc-MFS
ENG:ANIMEAT	0.72
ENG:ARTINFO	0.61
ENG:CONTCONT	0.71
ENG:LOCORG	0.54
ENG:PROCRES	0.69
DA:CONTCONT	0.66
DA:LOCORG	0.64
SPA:CONTCONT	0.58
SPA:LOCORG	0.63

Table 7.1: Accuracy for the MFS baseline

Table 7.1 shows the MFS baseline accuracy for each dataset. Not every dataset has the same sense distribution, and thus, MFS is not as hard a baseline for each. For instance, ENG:LOCORG has a much lower MFS (0.54) than ENG:ANIMEAT (0.72).

Moreover, not every lemma in each dataset would have the same MFS baseline if we implemented lemma-wise WSD instead of class-wise WSD, thus evaluation each lemma independently instead of all the lemmas of a dot type pooled together.

As an illustration of the irregularity of lemma-wise sense distributions, figure 7.1 shows the proportion of non-literal (metonymic plus underspecified) examples for the LOCATION•ORGANIZATION words that are common across the three languages in the datasets: *Afghanistan*, *Africa*, *America*, *China*, *England*, *Europe*, *Germany*, and *London*.

The values of the histogram are thus the complementary values to the most frequent sense, which is literal in these datasets. We can see that individual words show sense skewness and deviate from the overall MFS for DA:LOCORG, ENG:LOCORG and SPA:LOCORG.

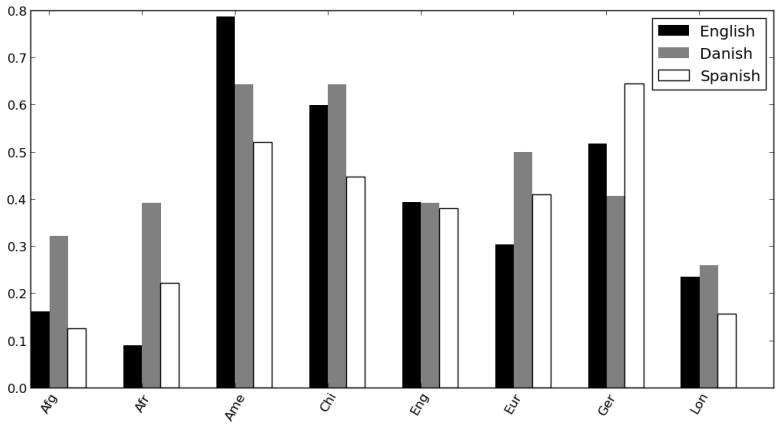

Figure 7.1: Proportion of non-literality in location names across languages

This skewness is a consequence of the way that each word is used in each 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 for all three languages, as it is often referred to as a placed that is acted upon.

7.3 Results

For all the combinations of dataset, annotation variants feature set, and training algorithm, we have trained and tested the system using ten-fold cross-validation with ten repetitions with random restarts. This method is commonly noted as 10×10 CV.

In this Section we analyze the results of running WSD on the nine datasets. The first parameter we abstract away from the analysis is **classifier choice**. Out of the three classification algorithms we have experimented with (Naive Bayes, decision trees and logistic regression), logistic regression does best for most datasets and feature sets in terms of overall accuracy. In this section we only refer to the performance of **logistic regression**. Appendix E provides the performance in terms of accuracy for all three classification algorithms.

Once we have settled for a learning algorithm, we need to settle for **a single feature set** to narrow down our analysis to a single system. Having fixed the learning algorithm to logistic regression also implies that the BOW baseline is a logistic regression classifier trained on the BOW feature set. Unlike BOW, the other baselines compared in Appendix D do not require a learning algorithm to be evaluated against MFS.

Table 7.2 shows the ranked top three feature sets for each dataset trained using logistic regression. Cf. Table 6.8 for an expansion of each feature sets into its constituting feature groups. We have obtained the ranking by sorting the different systems according to their overall accuracy.

Dataset	VOTE
ENG:ANIMEAT	alb, **all**, sem
ENG:ARTINFO	plb, **all**, alb
ENG:CONTCONT	pbr, pbw, alb
ENG:LOCORG	plb, alb, ple
ENG:PROCRES	pbr, pbw, alb
DA:CONTCONT	pbr, pbw, wnt
DA:LOCORG	**all**, alb, plb
SPA:CONTCONT	alb, **all**, plb
SPA:LOCORG	**all**, alb, pbw

Table 7.2: Feature set performance ranking

We choose the ALL feature set as the reference feature set for the quantitative analysis because out of the nine datasets, it is most often the best-scoring dataset. When not the best, ALL is either second or third, with few exceptions. For ENG:CONTCONT, `all` is the four dataset out of fourteen, performing at an overall accuracy of 0.70, whereas `alb`—the third feature set—yields 0.71.

The ENG:PROCRES dataset is different in that the highest-scoring feature sets

include PLE, which does not contain any semantic feature groups (cf. Section 6.2). The two feature sets that outperform PLE are PBR and PBW, which instead of using the lemmas of heads and dependents as features use the Brown-clustered heads and dependent from the feature group `bh` and `bh`.

Other than the ALL feature set, the preference of the datasets annotated with VOTE leans towards the feature sets PLB and PBR. In PLB we find an extension of the strictly grammatical features that use the lemmas of heads and dependents (as in the feature set PLE) with a bag of words. This feature set is not the most heavily engineered because it does not use any of the clustered semantic features or an LKB (cf. Table 6.8). The PBR dataset replaces the lemmas of heads and dependents with their Brown-cluster index, thus projecting the list of lemmas into a smaller amount of clusters, which improves recall.

Besides ALL and ALB, which only differ in the later also incorporating a bag of words, the schemes that rely on the Brown-clustered heads and dependents fare well in general. This indicates that using Brown clusters is a good compromise between using (possibly too sparse) lexical relations and using (possibly too coarse) POS tags. The ALB feature set is an extension of ALL with a bag of words, which makes it better that ALL in some instances but is much larger, therefore we use ALL as **reference feature set**.

Table 7.3 lists the accuracy obtained by training and testing the VOTE variant of each dataset on the `all` feature set (cf. Table 6.7) using logistic regression.

Moreover, the table also provides the error reduction over the MFS and BOW baselines. *Error reduction* (ER) is defined as the difference of error between systems (s_1, s_2), over the error of the original system (s_1), expressed in percentage. *Error* is defined as $1 - Accuracy$:

$$ER(s_1, s_2) = \frac{Error_1 - Error_2}{Error_1} = \frac{(1 - Acc_1) - (1 - Acc_2)}{1 - Acc_1} \tag{7.1}$$

The cases for the ALL and for BOW that are outperformed by MFS are marked with a dagger (†).

	Acc-ALL	Acc-MFS	ER-MFS	Acc-BOW	ER-BOW
ENG:ANIMEAT	0.84	0.72	42.86%	0.81	15.80%
ENG:ARTINFO	0.66	0.61	12.82%	0.61	12.82%
ENG:CONTCONT	0.83	0.71	41.38%	0.74	34.62%
ENG:LOCORG	0.73	0.54	41.30%	0.63	27.03%
ENG:PROCRES	0.6†	0.60	-29.03%	0.58†	4.76%
DA:CONTCONT	0.58†	0.66	-23.53%	0.61†	-7.69%
DA:LOCORG	0.68	0.64	11.11%	0.64	11.11%
SPA:CONTCONT	0.73	0.58	35.71%	0.7	10%
SPA:LOCORG	0.75	0.63	32.43%	0.64	30.56%

Table 7.3: Accuracies and error reduction over MFS and BOW for VOTE

All the accuracy scores for statistically significant over their respective MFS except those marked with a dagger. However, overall accuracy blurs out the details of the performance over each sense.

Table 7.4 provides the sense-wise F1 scores for VOTE for the reference feature set (ALL) and learning algorithm (logistic regression).

Dataset	L	M	U
ENG:ANIMEAT	0.88	0.68	0.00
ENG:ARTINFO	0.54	0.77	0.02
ENG:CONTCONT	0.89	0.60	0.00
ENG:LOCORG	0.79	0.61	0.00
ENG:PROCRES	0.45	0.71	0.01
DA:CONTCONT	0.73	0.14	0.09
DA:LOCORG	0.82	0.49	0.14
SPA:CONTCONT	0.81	0.64	0.17
SPA:LOCORG	0.85	0.68	0.02

Table 7.4: Sense-wise performance in terms of F1

Dataset	p	r	F1
ENG:ANIMEAT	0.00	0.00	0.00
ENG:ARTINFO	0.04	0.01	0.02
ENG:CONTCONT	0.00	0.0	0.00
ENG:LOCORG	0.00	0.00	0.00
ENG:PROCRES	0.03	0.00	0.01
DA:CONTCONT	0.14	0.08	0.09
DA:LOCORG	0.2	0.12	0.14
SPA:CONTCONT	0.27	0.15	0.17
SPA:LOCORG	0.05	0.01	0.02

Table 7.5: Performance for underspecified in precision, recall and F1

For this training data, feature set and learning algorithms, the results for the alternating senses (literal and metonymic) can reach high F1 scores. However, the classifier does poorly at identifying the underspecified sense. This lower performance suggests that the mapping between features and sense is less defined.

The underspecified sense is in general the least frequent sense for most datasets, which means there is less training data to identify it using the single classifier we have described in 7.1.

We have been able to develop a WSD system that captures the literal and metonymic senses with good F1 scores, but the underspecified sense remains difficult to identify.

7.4 Classifier ensemble

In Section 7.3 we have seen that the results for a three-way WSD using ALL on logistic regression yielded good results for the alternating senses and bad results for the underspecified sense. This indicates the postulation of the underspecified sense as one third, isolated sense is an unwieldy way of representing regular

polysemy at the token level. In Boleda et al. (2008), we find a recommendation to model the intermediate cases of regular polysemy not as separate classes but as the overlap of their constituting classes. With this remark in mind, we use a classifier ensemble to try to identify the underspecified sense.

In Sections 2.4 and 2.5.2 we have described the underspecified sense as having either the properties of both sense, or of none. We can implement this understanding by training a binary classifier for the literal and one for the metonymic sense, and assign the underspecified to the examples that are either tagged as simultaneously literal and metonymic, or simultaneously non-literal and non-metonymic.

In Martínez Alonso et al. (2012) we implemented a classifier ensemble using fewer and simpler features and training on decision trees and K-nearest neighbor classifiers. Diversity is a key requisite for any classifier ensemble (Marsland, 2011, p. 162), that is, for the aggregation of the criteria of more than one classifier to be fruitful, each classifier needs to have a different bias. In the ensemble described in this chapter, we achieve diversity by training on variants of the datasets, but we keep the feature set and learning algorithm fixed.

7.4.1 Procedure

This section describes the steps we have used to implement the classifier ensemble.

1. For each dataset D with a sense assignment, we generate a dataset D_l where all non-literal examples (metonymic and underspecified) are marked as members of the negative class, and a D_m dataset where the non-metonymic examples are marked as negative.

2. We train a logistic regression classifier C_l on D_l and another classifier C_m on D_m. These two binary classifiers are strictly trained with the literal and metonymic as the positive class, while the other two respective senses are pooled into the negative class. This results in the underspecified sense never being used as positive class for any classifier.

3. When assigning senses to examples, we combine the output of C_l and C_m in a method similar to the VOTE SAM. Underspecified senses are assigned by logical equality (i.e. an XNOR operation): if both classifiers assign the positive class or both assign the negative class, we assign the underspecified sense. In this way we assign the underspecified sense when the example is marked as both literal and metonymic, or when it is marked as neither.

The system can also be implemented with a complementary change of representation by including the underspecified sense as positive examples for both classes in the training step. The output is insignificantly different, but the proportion of underspecified senses assigned when C_l and C_m agree on the positive or negative class is inverted.

7.4.2 Results

The ensemble system behaves differently than the single-classifier scheme in several ways. Table 7.6 shows the accuracy scores for C_l and C_m. The accuracy

values are high (> 0.70) for most classifiers, although some datasets do not yield accuracies above 0.70, with the literal classifier C_l for DA:CONTCONT faring the poorest.

Dataset	Acc-C_l	Acc-C_m
ENG:ANIMEAT	0.83	0.84
ENG:ARTINFO	0.76	0.68
ENG:CONTCONT	0.83	0.86
ENG:LOCORG	0.75	0.74
ENG:PROCRES	0.70	0.61
DA:CONTCONT	0.58	0.80
DA:LOCORG	0.75	0.84
SPA:CONTCONT	0.74	0.83
SPA:LOCORG	0.80	0.83

Table 7.6: Individual accuracies for C_l and C_m

The accuracy values for the individual classifiers are only an indication of how well the ensemble-internal classifiers work. We examine the sense-wise F1 scores of the resulting ensemble to determine whether the system captures the sense distinctions we are after. Table 7.7 shows the sense-wise F1 score for each dataset.

Dataset	L	M	U
ENG:ANIMEAT	0.88	0.64	0.01
ENG:ARTINFO	0.47	0.75	0.15
ENG:CONTCONT	0.88	0.66	0.04
ENG:LOCORG	0.79	0.58	0.07
ENG:PROCRES	0.41	0.67	0.12
DA:CONTCONT	0.69	0.11	0.20
DA:LOCORG	0.81	0.45	0.25
SPA:CONTCONT	0.78	0.62	0.25
SPA:LOCORG	0.85	0.64	0.17

Table 7.7: Sense-wise F1 scores for the ensemble system

The ensemble system we describe in Martínez Alonso et al. (2012) had a bias for the most frequent sense that was stronger than the bias shown by its individual classifiers. In this ensemble, the bias for the most frequent sense is not as pervasive. Figures 7.2 and 7.3 show the difference between single and ensemble classifiers for the literal and metonymic sense respectively.

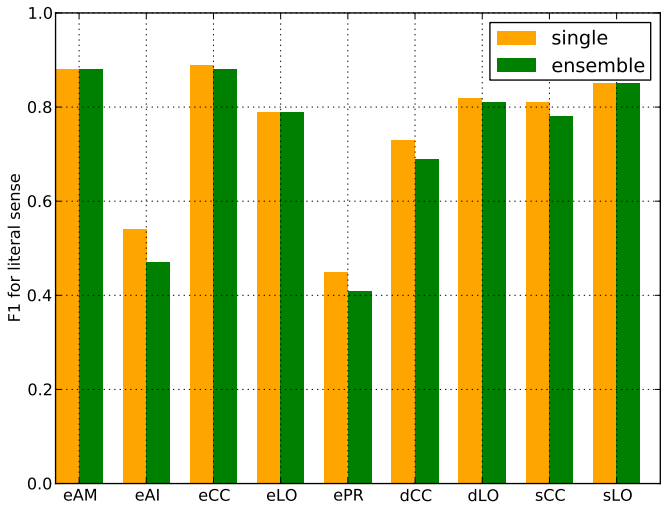

Figure 7.2: F1 for the literal sense in VOTE

The single and ensemble classifiers obtain very similar scores for the literal sense, although the single classifier is systematically better.

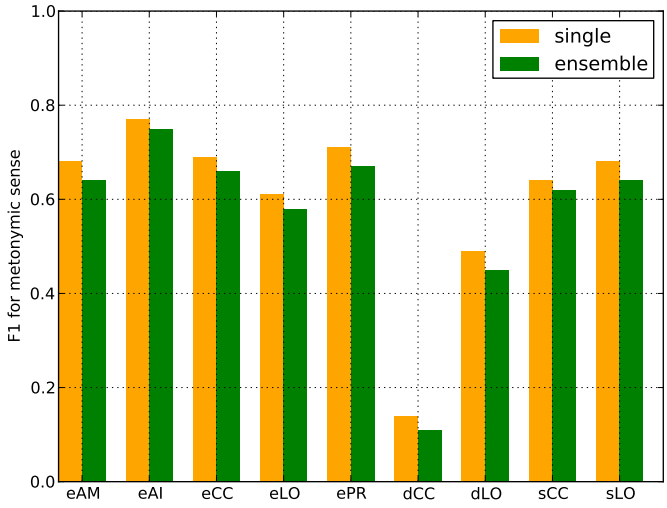

Figure 7.3: F1 for the metonymic sense in VOTE

The relative behavior of the systems is different for the metonymic sense, in that the drop in the ensemble system is even larger. The ensemble system is similar to—but still worse than—the single classifier.

With regards to the literal and metonymic senses, the ensemble does not

contribute in a favorable way. With regards to the underspecified sense, the ensemble system improves over the single classifier. However, the F1 scores are still very low, and neither the precision nor the recall of the system for the underspecified sense indicate that this approach is useful to capture the underspecified sense. Table 7.7 shows the sense-wise F1 score for the ensemble.

Dataset	p	r	F1
ENG:ANIMEAT	0.18	0.24	0.01
ENG:ARTINFO	0.24	0.30	0.15
ENG:CONTCONT	0.01	0.02	0.04
ENG:LOCORG	0.12	0.21	0.07
ENG:PROCRES	0.03	0.05	0.12
DA:CONTCONT	0.05	0.15	0.20
DA:LOCORG	0.10	0.20	0.25
SPA:CONTCONT	0.21	0.36	0.25
SPA:LOCORG	0.16	0.22	0.17

Table 7.8: Precision, recall and F1 for underspecified

Underspecified senses are assigned by the ensemble when the predictions of C_l and C_m coincide, either being positive or both negative. Table 7.9 lists the amount of underspecified senses in the test data for each of the ten runs. The columns describe the amount of expected underspecified examples in the test data for each fold (E), the amount of predicted underspecified examples by the ensemble each fold (U/r), and the amount of those predicted examples that are assigned by exclusion (Ex/f), namely when both C_l and C_m consider an example to be of the negative class.

The system overestimates the amount of underspecified examples, which damages the precision of the underspecified sense. However, this overestimation of the underspecified sense does not aid recall. Most of the underspecified senses are assigned by exclusion. This ratio would is inverted if the system is trained

Dataset	E/f	U/f	Ex/f
ENG:ANIMEAT	0.7	1.85	1.33
ENG:ARTINFO	5.4	9.67	8.79
ENG:CONTCONT	2.5	3.54	3.02
ENG:LOCORG	2.2	5.07	3.96
ENG:PROCRES	4.8	9.42	8.11
DA:CONTCONT	9.1	11.6	10.9
DA:LOCORG	8.3	10.62	10.11
SPA:CONTCONT	6.9	11.51	10.54
SPA:LOCORG	4.7	6.63	5.83

Table 7.9: Run-wise amount of underspecified senses and amount of underspecified assigned by exclusion

by considering the underspecified senses as positive examples in D_l and D_m, although such change of definition has no significant effect on the F1 scores for any sense in this setup.

7.5 Conclusions

In Section 7.1 we described the method for a WSD task to predict the annotated sense for our dot-type datasets. We have established that the system fares best using logistic regression as learning algorithm, and the ALL feature set to represent the linguistic information.

The system obtains very low scores for the recognition of the underspecified sense. In Section 7.4, we have devised a classifier ensemble made of one binary classifier C_l for the literal sense and one for C_m the metonymic sense. The ensemble assigns the underspecified sense when both classifiers agree in their prediction.

We have been able to predict the alternating class-wise literal and metonymic senses with acceptable performance using a single-classifier setup. The F1 scores are as high as 0.88 for literal and 0.77 for metonymic, but the performance for the underspecified sense was very low, with F1 scores of at most 0.17.

As an alternative to modelling the underspecified sense as a separate category, we have used a 2-classifier ensemble. One classifier was trained to identify the literal sense, and the other classifier was trained to identify the metonymic sense.

We have defined the underspecified sense as encompassing the exclusion and the intersection of literal and metonymic senses. Thus, the predicted underspecified senses for this ensemble system are those where both classifiers assign the positive class or the negative class at the same time (the logical equality of the output of both classifiers).

The classifier ensemble drops in performance for the literal and metonymic, and it improves for the underspecified. However, the underspecified sense does not receive an F1 better than 0.25 for any dataset.

Thus, the single-classifier WSD experiments do not justify postulating an independent, underspecified sense at the token level, and the classifier-ensemble experiments do not justify representing the underspecified sense as the logical equality of literal and metonymic.

Chapter 8

Literality prediction

We have used WSI (cf. Chapter 5), and single-classifier and 2-classifier WSD (cf. Chapter 7) to attempt to identify the underspecified sense as a discrete category. The results have been thus far negative, in that the systems have not found empirical evidence to include a third, independent underspecified sense in our sense inventories for token-wise dot-type representations.

In addition to these three different methods to determine the empirical validity of the underspecified sense, we analyze the advantages of a representation of the senses of a dot type in a continuum, where the underspecified sense is only implicitly represented as an intermediate value in the gradient between literal and metonymic.

In this chapter, we extend WSD to model the literality of examples in a continuum. In this way, a completely literal example would have a literality score of 1, a fully metonymic one would have a 0, and an underspecified example a value around 0.5. These values are a continuous projection of the literal-underspecified-underspecified sense gradient without any assumption of sense discreetness

The goal of this experiment is to test the adequacy of a continuous representation of the literal, metonymic and underspecified senses in a continuum. This understanding of the literality of an example as a continuous value leads to a redefinition of the WSD task into a task where we attempt to predict the literality score of examples based on their linguistic features. Predicting a numeric value in a supervised manner is called *regression*.

Finding a continuous account for the senses of dot types (without a discrete underspecified sense) that can be modelled with a method similar to our word-sense disambiguation experiments would provide an alternative representation for the dot type at the token level.

The interesting of predicting the literality value of an example is mainly for modelling purposes. Regression models allow the quantification of the explanatory power of features with regards to the dependent variable, and are thus useful for the assessment the learnability of a linguistic phenomenon.

Nevertheless, WSD is an NLP task that is not an application unto itself either, but is rather a preprocessing to aid other tasks like machine translation, information retrieval or question answering (Navigli, 2009). Thus, a continuous representation of the dependent variable of a sense-identification task is potentially more useful as a feature than a discrete value, provided that such

continuous value is a suitable representation of the semantic phenomenon at hand.

In Section 8.1 we describe how we calculate the continuous representation of literality. In Section 8.2 we describe the metrics used to evaluate regression, and provide the evaluation for the literality prediction system. In Section 8.3 we provide an analysis of the features—in particular, the grammatical features—that receive a larger coefficient and are more important to predict literality. In Section 8.4 we offer a comparison between the WSD system in Chapter 7 and the literality-prediction system in this chapter. Finally, 8.5 sums up the conclusions of this chapter.

8.1 Method

In this section we describe how we carry out the regression experiments. We use the features from the ALL feature set (cf. Table 6.8) as explanatory variables and we define a literality score LS from the annotations for each example as dependent variable. Instead of using the senses assigned with VOTE or MACE, we obtain the LS from the raw counts of the total annotations.

For each dataset D, we generate a dataset D_{LS} where the sense annotations are replaced by a literality score LS. We define the *literality score* of an example as the proportion of annotations that deem it literal. Each item receives a literality score LS. If an item has a set of annotation A, for each a_i from A:

a) if $a_i = literal$, add 1 to LS
b) if $a_i = metonymic$, add 0 to LS
c) if $a_i = underspecified$, add 0.5 to LS

When all the a_i values have been added, we divide LS by $|A|$ (the number of annotators for an item), thus normalizing LS to a value between 0 and 1.

Notice that, even though the values for literal (1.0) and metonymic (0.0) are valid by definition because we describe them as the poles of the literality continuum, the 0.5 for the underspecified sense is based on our understanding of it as a point between the literal and the metonymic.

Nevertheless, we consider this assumption reasonable. If an example were annotated by four annotators, twice as literal and twice as metonymic ($A = [L, L, M, M]$), the LS for this example would be 0.5 ($\frac{1+1+0+0}{4} = 0.5$), which is also the LS for an example annotated with four underspecified senses. That is, the LS for total disagreement between literal and metonymic is the same LS as for total agreement on the underspecified sense.

Moreover, if the arguments from Section 2.3 about a continuum between literality and figurativeness hold, a reasonable position for the underspecified values is between the two poles of the gradient. Notice that by defining two poles, we consider regular polysemy to be bound between a maximum and a minimum value of literality. Note that this idea of "minimum of literality" does not necessarily hold for all figurative-sense phenomena (cf. Section 2.3).

After generating the datasets D_{LS} with literality scores with the ALL feature set, we train and test a regression model using Bayesian Ridge (MacKay, 1992) as implemented in SkLearn (Pedregosa et al., 2011). Bayesian Ridge regression is time-consuming but it fits its parameters on the data at hand and requires no explicit parametrization. Moreover, Bayesian Ridge is more robust towards

noisy data than other methods like Ordinary Least Squares regression. The system is trained and tested using 10×10 CV.

We have settled for the ALL feature set as reference feature set in Chapter 7, and we do not experiment with other feature group combinations. The only feature group that is left out of the regression model is the bag of words `bow`. However, the features in ALL incorporate Brown-clustered lexical information in the `bb` feature group.

Figures 8.1 and 8.2 show the distribution of the literality score *LS* in the gold standard for English, Danish and Spanish. The lines are smoother the more annotators a dataset has.

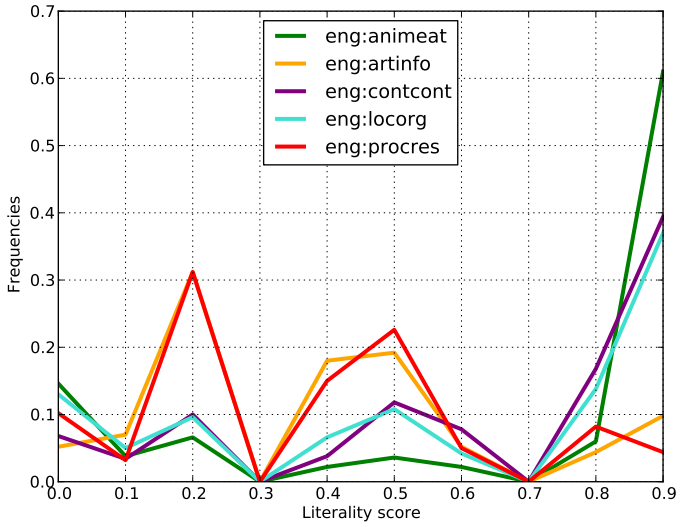

Figure 8.1: LS distribution for English

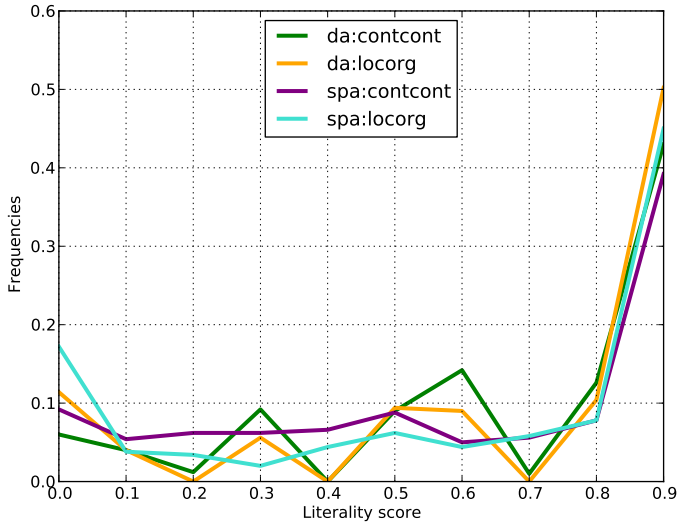

Figure 8.2: LS distribution for Danish and Spanish

Even though LS is a continuous variable, the number of different values we can generate is limited by the amount of annotators. Indeed, the unpopulated values of LS are an artifact of the number of the annotators: the different combinations of literal, metonymic and underspecified votes for the five annotators of the English dataset never provide a value of 0.3 or 0.7. Notice how there are no values for 0.2, 0.4 or 0.7 for Danish. Danish has fewer annotators, and therefore fewer possible intermediate values for LS. Spanish, which has the most annotators, presents smoother curves.

8.2 Evaluation

8.2.1 Evaluation metrics

There are several metrics to evaluate the goodness of fit of a regression model. We use two of them, the *mean squared error* (MSE) and the *coefficient of determination* (R^2), to evaluate the performance of our regression experiment.

MSE is a loss function defined as

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2. \tag{8.1}$$

Where y are the expected values of the dependent variable, and $\hat{y}$ are the predicted values of the dependent variable. Thus, MSE is the mean of the square of differences between the predicted and the expected values. Since MSE measures the difference between expected and predicted values, lower values of MSE will indicate a closer fit and are more desirable. MSE has a lower bound at zero, but it has no formal upper bound, i.e. scaling y by two would also scale the values for MSE accordingly. This lack of an upper bound makes MSE to be only interpretable with regards to a baseline.

R^2 is a metric that, instead of measuring error rate, aims at estimating how well future examples can be predicted by the model. The definition of R^2 is:

$$R^2(y, \hat{y}) = 1 - \frac{MSE(y, \hat{y})}{MSE(y, \overline{y})} \tag{8.2}$$

Where y are the expected values of the dependent variable, $\hat{y}$ are the predicted values of the dependent variable, and $\overline{y}$ are the averaged values of y. Thus, R^2 is the ratio of the MSE between the expected and predicted, and the MSE between the expected and the mean values of the dependent variable. The quotient is what makes R^2 a coefficient to measure the capacity of the model to generalize over the observed data.

The values of R^2 are normalized, and usually defined between 0 and 1, although there can be regression algorithms that yield negative values. The higher the value of R^2, the more the explanatory variables (in our case, the features) are accountable for the value of the dependent variable (in our case, the literality score LS). R^2 has a more direct interpretation as the proportion of the variance of the dependent variable that the regression algorithm can explain from the explanatory variables.

8.2.2 Results

Table 8.1 shows the results for the literality prediction models. MSE-$\overline{LS}$ is the MSE obtained by assigning each example with the average literality score ($\overline{LS}$), MSE-BR is the MSE for the system using Bayesian Ridge regression, and R^2 is the coefficient of determination for Bayesian Ridge; the R^2 from the $\overline{LS}$ baseline offers no assessment on the informativeness of the features to predict the dependent variable LS and we do not provide it. All the nine datasets improve significantly over the $\overline{LS}$ baseline on a corrected paired t-test with $p < 0.05$.

	MSE-$\overline{LS}$	MSE-BR	R^2
ENG:ANIMEAT	0.16	0.10	0.37
ENG:ARTINFO	0.07	0.06	0.14
ENG:CONTCONT	0.10	0.07	0.33
ENG:LOCORG	0.14	0.01	0.30
ENG:PROCRES	0.07	0.06	0.11
DA:CONTCONT	0.10	0.07	0.28
DA:LOCORG	0.13	0.08	0.33
SPA:CONTCONT	0.12	0.08	0.26
SPA:LOCORG	0.14	0.07	0.49

Table 8.1: Evaluation for literality prediction

In a manner similar to the WSD experiments from Chapter 7, each dataset poses a different learning difficulty. For the English datasets, literality prediction for ENG:LOCORG fares much better than for ENG:PROCRES. The later dataset has very low α and the values of the annotations—and thus of LS—are not very reliable[1], which directly impacts how much of a mapping a learning algorithm

[1]note that we qualify the data using the reliability coefficient, because the only different between different annotation subtasks is the data itself

can establish between the features and the dependent variable.

In this section's scatter plots, the black straight line represents the ideal behavior of the system, where $\hat{y}_i = y_i$. The green dots represent data points that receive a $\hat{y}_i$ that is within $0.95y_i > \hat{y}_i < 1.05y_i$ and represent good approximations. The points in red are those where $|\hat{y}_i - y_i| > 0.5$, and represent gross mispredictions of the value of LS. Yellow points represent formally illegal values of LS, that is, $\hat{y}$ values above 1.0 or below 0.0, which do not appear in the training data. The total nine scatter plots, one for each dataset, are provided in Appendix F.1.

On one hand, the scatter plot for ENG:LOCORG shows more red points that represent values that are mispredicted, and yellow points that that are outside of the expected [0,1] interval for literality. Its MSE-BR is the lowest of all datasets, but the values of MSE are not strictly comparable across datasets. On the other hand, the predictions for ENG:PROCRES deviate less from the mean. For this dataset, in spite of the smaller amount of values marked in red and yellow, the goodness of fit of the model is lower because the predicted values of LS are closer to each other.

The graphical representation of the predictions shows that the model for ENG:LOCORG approximates the ideal $\hat{y}_i = y_i$ line better than ENG:PROCRES. Also, the R^2 value for ENG:LOCORG is about three times larger than for ENG:PROCRES, which we interpret as a consequence of the linguistic features being three times more informative as explanatory variables to determine the LS of the former dataset.

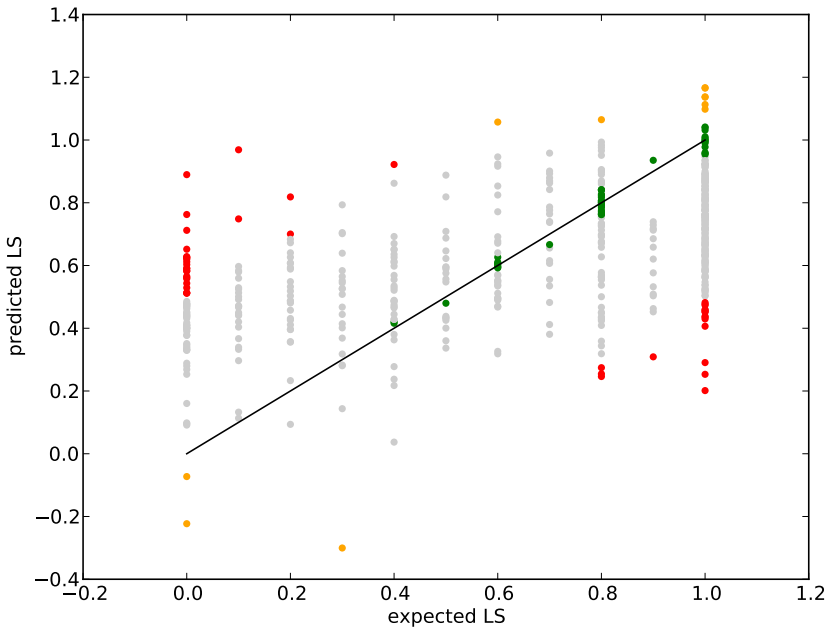

Figure 8.3: Literality prediction for ENG:LOCORG

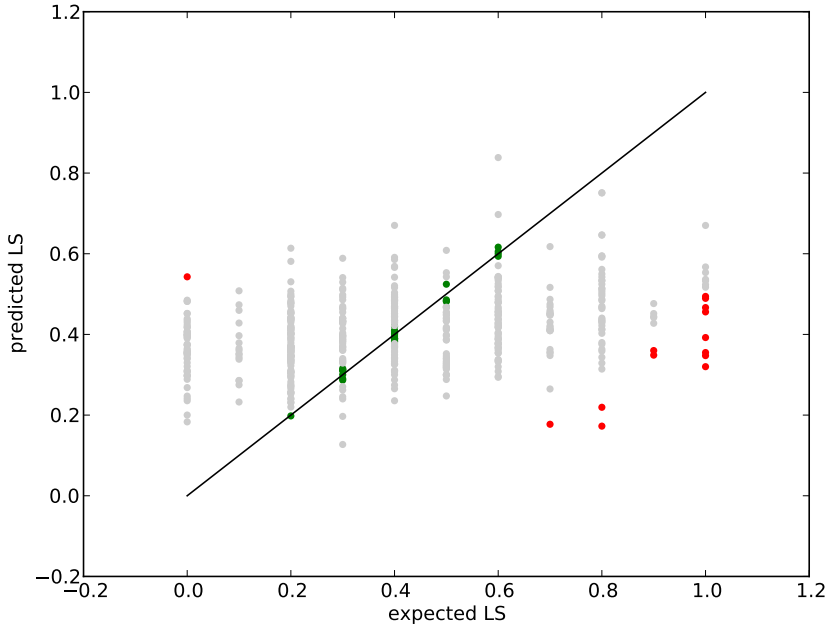

Figure 8.4: Literality prediction for ENG:PROCRES

There is a difference in granularity of the values of the expected LS across datasets. This difference is a result of the different amount of annotators for each language. Each of the $\binom{|A|}{3}$ possible combinations of annotations (where $|A|$ is the number of annotators and 3 is the number of possible senses) yields a particular value of LS, and even though it is an injective operation (we can calculate LS from A, but cannot know A from LS because some combinations obtain the same LS), the bigger the set A, the more fine-grained the literality score values of LS will be.

Figures 8.5 and 8.6 show this difference in granularity. The values for the expected LS in DK:LOCORG are more spaced between them than the values for SPA:LOCORG because most Danish examples are annotated by three volunteers, whereas most of the Spanish examples are annotated by six.

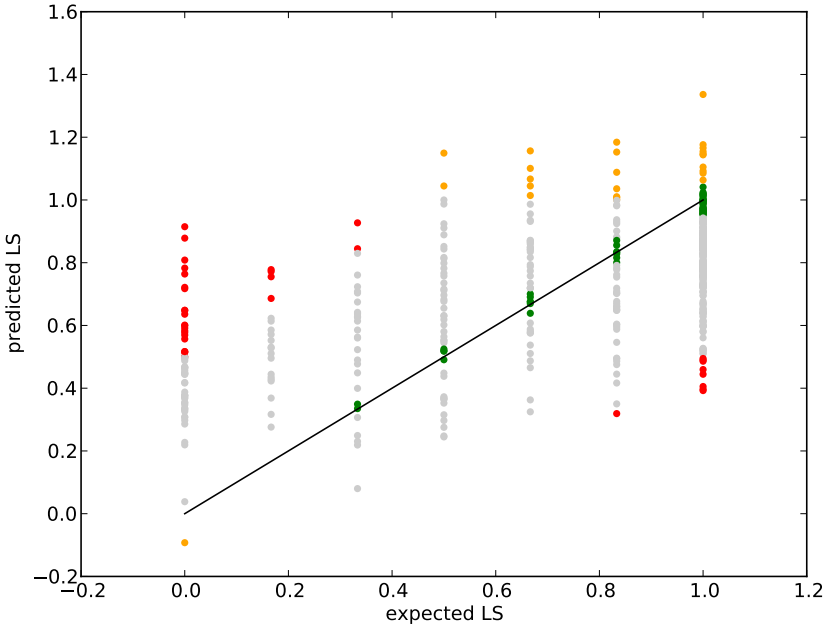

Figure 8.5: Literality prediction for DA:LOCORG

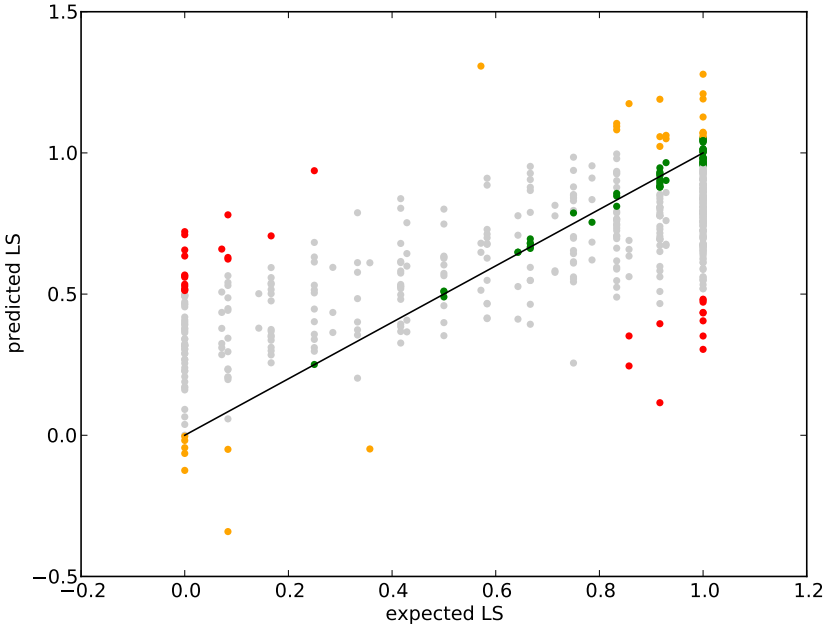

Figure 8.6: Literality prediction for SPA:LOCORG

The low-agreement datasets ENG:ARTINFO and ENG:PROCRES are the hardest to fit. This is a result of the sense distinctions being more difficult to annotate, which makes the dependent variable noisier, and the ability of the regression method to find reliable coefficients for the linguistic features less likely. This lower performance is made apparent by the low R^2 values (0.14 and 0.11) for these datasets.

A_o	α	R^2
ENG:ANIMEAT	ENG:ANIMEAT	**SPA:LOCORG**
DA:LOCORG	**SPA:LOCORG**	ENG:ANIMEAT
ENG:LOCORG	DA:LOCORG	**ENG:CONTCONT**
ENG:CONTCONT	ENG:LOCORG	DA:LOCORG
SPA:LOCORG	SPA:CONTCONT	ENG:LOCORG
ENG:PROCRES	DA:CONTCONT	DA:CONTCONT
ENG:ARTINFO	**ENG:CONTCONT**	SPA:CONTCONT
SPA:CONTCONT	ENG:ARTINFO	ENG:ARTINFO
DA:CONTCONT	ENG:PROCRES	ENG:PROCRES

Table 8.2: Ranking of the nine datasets according to their A_o, α and R^2.

Table 8.2 shows a ranking of the nine datasets according to their A_o, α and R^2. The determination coefficient R^2 is much more closely paired with α than with A_o. This indicates that, even though α is strictly a measure of the replicability of an annotation procedure, it provides a better estimate of the validity of a dataset than observed agreement alone. However, R^2 is not completely correlated with α. Besides exchanging the first and second role between ENG:ANIMEAT and SPA:LOCORG, the ENG:CONTCONT dataset behaves better than expected at literality regression.

ENG:CONTCONT does not have very high α (0.31), but we believe that it is precisely the variance in its annotation that makes this dataset suitable for the task at hand. In the scatter plot for ENG:CONTCONT (cf. Figure F.5), the area in the middle of the plot, which corresponds to the middle third of literality values—those that correspond to the underspecified sense in the VOTE SAM—is more populated than the area for those similar values in ENG:ANIMEAT (cf. Figure F.3).

The A_o of ENG:ANIMEAT and ENG:CONTCONT is respectively 0.86 and 0.65. The lack of mid-literality data for ENG:ANIMEAT, which has a very high A_o, penalizes its performance in the middle third of LS, whereas ENG:CONTCONT has more evenly distributed LS values and the regression model performs better than what the α ranking suggests.

We have successfully fit a Bayesian Ridge regression model on each dataset, using the ALL feature set as explanatory variables and the literality score LS as dependent variable.

8.3 Feature analysis

Fitting a regression models yields a coefficient for each feature. Coefficients can be negative or positive, thus helping to decrease or increase the value of

the predicted variable. We can also examine the coefficients for each of the nine regression models. Appendix F.1 provides the top 20 positive and top 20 negative coefficients for each dataset. We have obtained these coefficients by fitting the regression algorithm on all the 500 examples for each dataset and not by 10×10 CV, in order to obtain one stable ranking of features calculated at one for each dataset.

We find that the features from the `bb` feature group are pervasive among the most important features for all datasets. This feature group is also the largest after the bag of words `bow`, which is not included in the ALL feature set. Grammatical features that correspond to frequent words or syntactic roles receive high coefficients. In the following sections we comment the features that the regression model uses to give low or high LS scores. We focus on the grammatical features. As in the previous chapters, h stands for the dot-type headword of each example.

8.3.1 English datasets

In ENG:ANIMEAT, in example a) we find that h being plural is the strongest indicator of literality. This supports the correspondence between this dot type and a mass/count alternation where the literal sense is a count noun. However, participating in a coordination gives low literality for b), but we consider this a corpus feature and not a general trait of the behavior of this class; in this corpus, enumerations of kinds of food are more common than enumerations of fauna.

In example c) (ENG:ARTINFO), as in the previous example, plural makes h more literal, because the ARTIFACT sense corresponds more to a count reading, and the INFORMATION to a mass reading. If h is headed by a preposition as in d) it is more likely to be metonymic.

In ENG:CONTCONT, for example e) the high-coefficient features `bc:10001110` and `bc:0001110111` stand for h being complemented by a word from the Brown clusters numbered 0001110111 or 10001110. These clusters contains words of substances that are used for making vessels, like *plastic* or *glass*, or adjectives like *red* and *wooden*, which normally describe objects. These features that describe how an object is made aid the literal reading, along the feature for plural.

In ENG:LOCORG, the feature `phw:in` (h begin a dependent of *in*) is the most relevant feature for a literal reading in f), but also the more general `pl:pmod`, that is, h being the nominal head of a prepositional phrase. This is symmetric with the preference of metonymic senses to be placed in the subject position as in g).

In ENG:PROCRES, the reading is more literal if h is an explicit argument as the object of the preposition *of* as in h), and the reading is more metonymic if h is the object of a verb as in i). In Section 2.6.5 for a remark on argument satisfaction being a relevant feature to disambiguate between eventive and resultative readings.

(8.3) a) The *rabbit**s*** were randomly divided into four groups of seven animals.

b) Grilled in foil or alongside a ham, *turkey* **or** chicken, those who shied away from onions before will delight in their new found

vegetable.

c) Holbrooke suffers from a strain of narcissism that impels him to quote himself, frequently and at length, including from diaries, articles, TV interviews, faxes, and private letters to the president.

d) Consider, too, that the quality of the paper, printing, and binding of such *dictionaries* is far superior to that of most other books available, creating a cost per copy of about $ 5.

e) Most of the regalia and artifacts on display, including the **red wooden** coffin *containers* in the burial vault, are copies of the items discovered during excavations in the 1950s .

f) Surrounded by elegant classical mansions and gilded wrought-iron grilles with ornamental gateways, which frame the marble fountains of Neptune and Amphitrite, Nancy's huge Place Stanislas is one of the most harmonious urban spaces in *Europe*

g) Last year, *Canada* ran a $ 21 billion trade surplus with the United States.

h) Woks and any other gadgets essential for Chinese cookery make good *purchases*.

i) Naturally, the lack of ready funds holds one back from vital *purchase* **of** an improved farm implement or even a fishing vessel.

8.3.2 Danish datasets

In DA:CONTCONT, if the headword h is introduced by a quantifier like *to* (*two*), its *LS* decreases. The feature for h being headed by *to* (`phw:to`) indicates structures like the one highlighted in example a). Cf. Section 6.1 for the particularities of the Danish treebank. Other features that provide low-*LS* interpretations are the prepositions *med* and *til* (*with* and *to*) as heads of h, as well as h being the head of a conjunction, having thus a subordinate clause.

The lexical head feature with a highest coefficient is `phw:i`, and it stands for h being headed by the preposition *i* (*in*) in b). Other grammatical features that aid a literal reading are h having a prepositional modifier that for instance indicates the material the container is made from.

In DA:LOCORG, there are fewer grammatical features with negative coefficients, but more with high positive coefficients. Again, examples where h is introduced by *i* are more often literal. The preposition *mod* (*against*) helps trigger a more metonymic reading when it is the head or the dependent of h as in example c).

(8.4) a) Hæld **to** *dåser* flåede tomater i gryden [...]
(Pour **two** *cans* of peeled tomatoes into the pot [...])

b) Deltagerne stikker benene **i** en *sæk* [...]
(The participants put their legs **in** a *sack* [...])

c) Først to kampe **mod** *Japan*, som man troede afgørende [...]
(First two matches **against** *Japan*, that were thought to be decisive [...])

8.3.3 Spanish datasets

In SPA:CONTCONT we find few grammatical features with a strong negative contribution to LS. However, the literal reading in a) is facilitated by the preposition **en** (*in*), and by *h* having an article.

In SPA:LOCORG the syntactic features are also only relevant to increase the literality of examples. Low-literality examples are identified by semantic features like wordnet types (`wn`) or brown-clustered bag of words (`bb`) features. Consistent with the other two languages, the preposition *en* is the strongest grammatical indicator for a high LS in c), along with *h* being the modifier of a noun, the word *Francia* in b) being the dependent of *París*.

(8.5) a) **El** primero de estos *botes*, rotulado con la leyenda 'veneno para osos', fue hallado el día 30 por un paseante.
(**The** first one of those *jars*, with the legend 'poison for bears' written on it, was found on the 30th by a stroller.)

b) La NSA escuchó las conversaciones telefónicas de la princesa sin la aprobación de los servicios secretos británicos, la noche de su fallecimiento en París (*Francia*) el 31 de agosto de 1997 [...]
(The NSA tapped the phone conversations of the Princess without the approval of the British secret services on the night of her demise in Paris (*France*) on the 31st of August 1997 [...])

c) Es cierto que no estuve ni **en** *Valencia* ni en Castellón [...]
(It is true that I was neither in Valencia nor in Castellón [...])

8.4 Comparing regression to classification

We have proposed a continuous representation for the senses of dot types as an alternative to the discrete three-way representation. After attempting to predict the discrete representation using WSD and the continuous representation with literality prediction, we want to compare the performance of both systems to determine which representation is more adequate.

To the best of our knowledge, there is no established metric for the comparison of performance between a regression and a classification system. In our case, we propose translating the error coefficient of our system into a metric with a value similar to accuracy, and thus, compare this metric with the accuracy for WSD.

If accuracy is the proportion of right classifier predictions, its complement (one minus accuracy) is the error rate. We can define regression accuracy (RA) as one minus the mean average error of the regression. Mean average error (MAE) differs from MSE in that the difference between expected and predicted value is expressed in absolute value and not squared. The values of MAE are larger than the values for MSE when y is defined between 0 and 1. Using MAE reduces the chance that we overestimate the performance of the regression system.

$$RA(y, \hat{y}) = 1 - \frac{1}{n}\sum_{i=1}^{n} |\hat{y_i} - y_i|. \tag{8.6}$$

In our case the value of *RA* is defined between 0 and 1 because the error rate is also defined between 1 and 0. This allow us to compare the values for RA and accuracy for each dataset. If the dependent variable *LS* were defined for instance between 1 and 5, we would not be immediately able to compare because the error-rate values would not be normalized.

Note that the error when counting accuracy is quantified discretely (hit or miss), whereas the error for regression is quantified as a difference between expected and predicted value and is a continuous value.

Table 8.3 shows the accuracy for the single classifier from Section 7.3 and the RA obtained from the regression experiments in this section.

Dataset	Acc	RA
ENG:ANIMEAT	**0.84**	0.76
ENG:ARTINFO	0.64	0.80
ENG:CONTCONT	**0.82**	0.79
ENG:LOCORG	0.73	0.75
ENG:PROCRES	0.61	0.80
DA:CONTCONT	0.57	0.79
DA:LOCORG	0.68	0.77
SPA:CONTCONT	0.69	0.76
SPA:LOCORG	0.75	0.79

Table 8.3: Comparison between accuracy and RA

The values of *RA* are more similar to each other than the accuracies for the different datasets. This suggests the continuous representations lends itself to more stable systems where the errors are not centered around an expected value for the dependent variable (the underspecified sense).

If we compare the difference between accuracy and *RA* strictly, we can see that the WSD for two datasets outperforms the literality prediction. These two datasets, ENG:ANIMEAT and ENG:CONTCONT present particularities with regards to the annotations for the underspecified sense (cf. Section 4.7.1).

The ENG:ANIMEAT dataset contains the fewest underspecified senses according to both the expert and the turkers, and in the ENG:CONTCONT dataset, all the underspecified senses are assigned by backoff.

The performance drops in these two datasets because there are fewer mid-literality values (cf. Table 8.1) and the dependent variable has less variance the system can give account for.

This comparison suggests that a continuous representation is a viable alternative for the discrete classification of senses in literal, metonymic and underspecified, provided that there are mid-literality values.

The absence of mid-literality values is caused by two factors, namely the turker annotator bias to disprefer the underspecified sense tag (cf. Section 4.9.1), and the absence of contexts that would receive an underspecified sense tag by volunteer or expert annotations.

The ENG:CONTCONT dataset exemplifies the first factor, because it has no example annotated as underspecified by plurality, but only by backoff. The ENG:ANIMEAT dataset exemplifies the second factor, because—in addition to the turker annotation bias—that words of the ANIMAL•MEAT dot type are sel-

dom copredicated. This dataset is also the one with the fewest underspecified examples in the expert annotation (cf. Table 4.3).

8.5 Conclusions

In this chapter we have introduced an alternative representation for the senses in a dot type as a continuous value between 0 (fully metonymic) to 1 (fully literal).

By examining the R^2 we determine that literality modelling with the linguistic features from Chapter 6 can give account for about one third of the variance of the literality score.

A desirable way to assess the appropriateness of this representation would be comparing the same model fitting against another regression model where the sense annotations are projected into a continuous value in another way, although we have not found a comparable representation (say, with the underspecified sense being less literal than the metonymic) sensible enough as a comparison scenario.

The system could be improved by setting all $\hat{y}_i > 1$ to 1, and $\hat{y}_i < 0$ to 0. This would be a postprocessing step that would improve MSE but would not allow us to claim a higher R^2. It would be an engineering degree but would have no impact on the coefficients we are obtaining, as they are calculated on training time, and would not improve our knowledge on the relation between explanatory (features) and dependent (LS) variables.

We successfully fit a regression model for each dataset using the literality score as a dependent variable. In a comparative study with WSD, we deem the literality score to be a more robust representation, and thus we consider this literality score LS an adequate candidate for the representation of token-wise dot-type meaning.

Chapter 9

Agreement prediction

In Chapter 4 we have covered the interpretation of Kripperdorff's α coefficient to describe the reliability of an annotation task, and the way that observed agreement (A_o) is calculated for each example. In this chapter we describe a method to predict the A_o of the examples in our datasets.

Krippendorff (2011) defines disagreement as *by chance*—caused by unavoidable inconsistencies in annotator behavior—and *systematic*—caused by properties of the data being annotated. We know some of the difficult examples in our dataset to have lower agreement, which can be a result of the linguistic characteristics of these examples.

Even though, strictly speaking, the value of α only provides an indication of the replicability of an annotation task, we suggest that the difficulty of annotating a particular example will influence its local observed agreement. Thus, easy examples will have a high A_o, that will drop as difficulty increases.

We have seen that lower-agreement examples are often marked as underspecified in the expert annotations, and we expect the result of this task to aid in the understanding of underspecification. Identifying low-agreement examples by their linguistic features would help characterize contexts that make dot-type words difficult to annotated and less likely to be interpreted as fully literal or metonymic.

The goal of this experiment is to measure how much of the disagreement in the annotations is caused by linguistic properties of the examples and is thus systematic.

We will consider the proportion of explained variance of the regression model described by the coefficient of determination R^2 to be the amount of disagreement our system can give account for, and is thus systematic.

9.1 Method

Observed agreement A_o is a continuous variable, and using supervised learning to predict it from the linguistic features in Chapter 6 is a regression task, similar to the regression task to predict literality in Chapter 7

For each dataset D, we generate a dataset D_{agr} where the sense annotations are replaced by the observed agreement A_o of each example. Note that the α coefficient is an aggregate measure that is obtained dataset-wise, and A_o is the

only agreement measure available for each individual example.

The method for agreement prediction is the same as in literality prediction (cf. Section 8.1), with the exception of the dependent variable. We use the ALL feature set as explanatory variables. It is still a class-based method because we are not looking at the identity of the headword for training. We train and text on Bayesian Ridge regression using 10×10 CV.

Notice that using A_o as dependent variable is different from using LS. The two variables are ideally orthogonal, that is, the observed agreement for an example with perfect agreement for the literal sense is 1, but so is the agreement for an example where all annotators provide the metonymic sense (9.1). However, the literality score LS for the fully literal example is 1, whereas the LS for a fully metonymic example is 0 (9.2).

$$A_o([L, L, L, L]) = 1 = A_o([M, M, M, M]) \tag{9.1}$$

$$LS([L, L, L, L]) = 1, LS([M, M, M, M]) = 0 \tag{9.2}$$

Figures 9.1 and 9.2 show the distribution of values of A_o for all datasets. The values for English fall in three ranges. The distribution of values is not less smooth than the values for LS in Chapter 8 (cf. Tables Figures 8.1 and 8.2) as a result of the way A_o is calculated from the pairwise agreements (cf. Section 4.6.1).

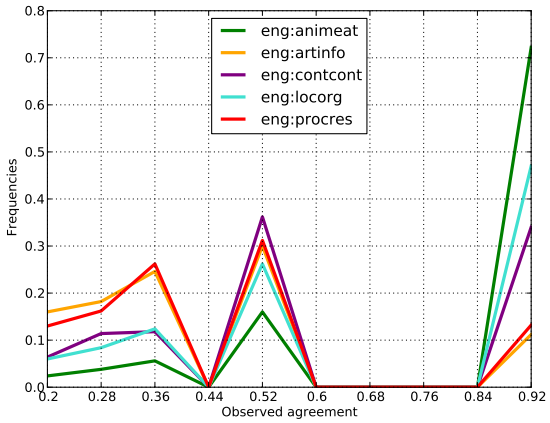

Figure 9.1: A_o distribution for English

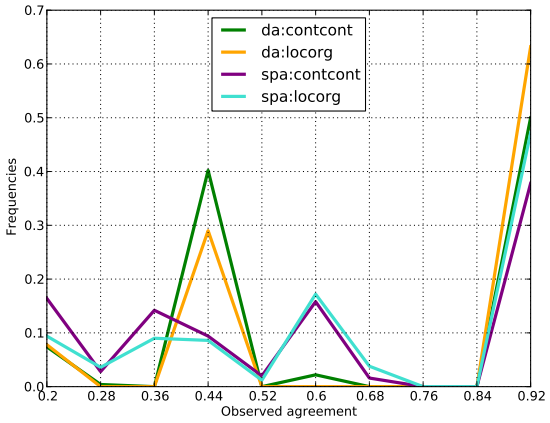

Figure 9.2: A_o distribution for Danish and Spanish

The values for Danish, with fewer annotators, are less smooth than for English, and the Spanish datasets have the smoothest distribution of A_o because they have the most annotators.

9.2 Results

Table 9.1 shows the scores for agreement prediction for all nine datasets. MSE-$\overline{LS}$ is the MSE obtained by assigning each example with the average observed agreement ($\overline{A_o}$), MSE-BR is the MSE for the system using Bayesian Ridge regression, and R^2 is the coefficient of determination for Bayesian Ridge. Cf. Section 8.1 for a review of the evaluation metrics for regression.

Dataset	MSE-$\overline{A_o}$	MSE-BR	R^2
ENG:ANIMEAT	0.06	0.06	-0.02
ENG:ARTINFO	0.06	0.05	-0.02
ENG:CONTCONT	0.08	**0.08**	**0.01**
ENG:LOCORG	0.08	**0.08**	**0.00**
ENG:PROCRES	0.06	0.06	-0.03
DA:CONTCONT	0.13	**0.13**	**0.05**
DA:LOCORG	0.13	**0.13**	**0.01**
SPA:CONTCONT	0.09	**0.09**	**0.03**
SPA:LOCORG	0.08	**0.07**	**0.02**

Table 9.1: Evaluation for agreement prediction

Notice there are negative values of R^2. This means that the system would be better approximated by disregarding all features and assigning the average A_o for these three datasets. For positive values of R^2, we can claim that there is at least that much proportion of the disagreement that can be explained by the features and is thus systematic.

Significant (corrected paired t-test with $p < 0.05$) improvements over the $\overline{A_o}$ baseline are marked in bold. The R^2 scores are lower for agreement prediction than for literality prediction (cf. Table 8.1), and the difference between MSE-$\overline{A_o}$ and MSE-BR can only be observed in the third or fourth decimal place, which do not appear in the tables.

The datasets that can be fit over baseline, and whose disagreement can be partially explained by the linguistic features are the datasets for CONTCONT and LOCORG for all three languages. Datasets with very high or very low agreement have too little variation for the system to be able to pick on patterns that relate the features to the dependent variable.

The Danish and Spanish datasets, annotated by volunteers, show more identifiable systematic disagreement, regardless of their α score. The dataset with more (5% for an R^2 of 0.05) identified systematic disagreement is DA:CONTCONT, which has been the most difficult dataset to automatically resolve with WSD, even disregarding the underspecified sense (cf. Section 7.3).

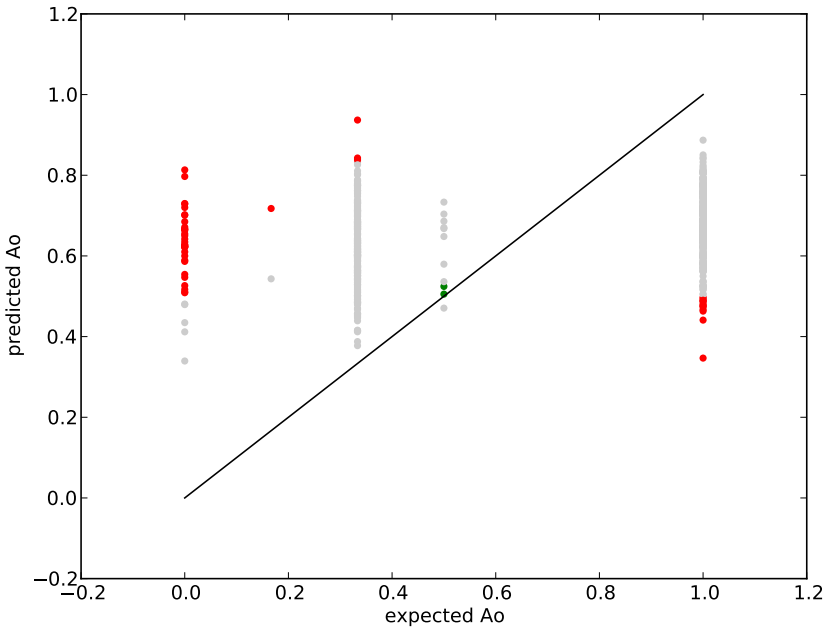

Figure 9.3: Agreement prediction scatter plot for DK:CONTCONT

Table 9.3 shows the scatter plot for DK:CONTCONT. The other plots are provided in Appendix G. We use the same color convention as in Section 8.2. Notice that there are no yellow points in the scatter plots for agreement prediction. Yellow points stand for formally invalid values (above 1.0 or below 0.0), which the system does not generate because there is less variance for the values of A_o than for the values of LS. This lack of variance can be seen in the less smooth values of A_o shown in Figures 9.1 and 9.2, which are grouped in ranges.

The datasets we cannot fit over baseline are ENG:ANIMEAT, ENG:ARTINFO, and ENG:PROCRES. The last two datasets have fared consistently badly in the

WSD experiments (cf. Section 5.3), and have low (0.12 and 0.10) α scores, besides being the two datasets with the highest difference (D) proportion in Table 4.12.

The high D proportion indicates that the annotations for these two datasets are not very stable, and the low α indicates that repeating the annotation task with the same examples would yield very different sense annotations. The impossibility to fit the agreement prediction system over the baseline suggests, along with the previous reasons, that these two datasets have no example-wise systematic disagreement. The inability of neither finding systematic disagreement or accurately identifying the senses of these datasets by WSD indicate that the annotations are not reliable.

The ENG:ANIMEAT dataset, which also resists agreement prediction, is different to ENG:ARTINFO and ENG:PROCRES in three aspects. Firstly, it has the highest α of all nine datasets. Secondly, it has the lowest amount of underspecified senses by either the expert or the turkers. Thirdly, it is the dataset with highest overall accuracy in WSD (cf. Section 7.3), and the second highest in literality prediction.

However, it is not possible for the system to identify any systematic disagreement in ENG:ANIMEAT. This is due to the low variance of the dependent variable, which makes this dataset the complementary to ENG:ARTINFO and ENG:PROCRES: the sense distinction is arguably too easy for the annotators, and the system cannot find a mapping between the linguistic features and the little disagreement.

9.3 Feature analysis

Fitting a regression model yields a coefficient for each feature. Coefficients can be negative or positive, thus helping to decrease or increase the value of the predicted variable. We examine the coefficients for the six datasets that fit over baseline. Appendix G provides the top 20 positive and top 20 negative coefficients for each dataset. We have obtained these coefficients by fitting the regression algorithm on all the 500 examples for each dataset and not by 10×10 CV, in order to obtain one stable ranking of features calculated at one for each dataset. As in the previous chapters, h stands for the dot-type headword of each example.

We observe that the grammatical features have positive coefficients more often than not. This indicates that explicit syntactic behavior helps the annotators take decisions that are consistent. Nevertheless, some few grammatical features correlate with a lower agreement, and we interpret them as causes of systematic disagreement. For the LOCORG datasets, we identify no grammatical features with high negative coefficients. For the CONTCONT datasets, however, we identify several grammatical features that help identify systematic disagreement.

When the preposition *in* is the head of h for an English CONTCONT word, the agreement decreases. This preposition is normally associated with literal meaning, but its presence also diminishes the agreement in examples like first English sentence in (9.3).

For Spanish, h being complemented by a quantifier is a feature that reveals possible low agreement. This is similar to the example d) in 4.8 for Spanish,

but also to the Danish example d) in 4.5. In both examples from Chapter 4 the container word is introduced by a quantifier, and the examples received minimal agreement.

For Danish, container nouns in definite form have low agreement. Although noun definiteness is a matter of morphology (cf. Section 2.6.6), this is an an example of systematic disagreement caused by pragmatics: the definite form is used when there is a referent that has already been mentioned previously, but the annotation has been carried out sentence-wise. Thus, many referents are not provided, and that makes the annotator's task more difficult.

Danish also has a specific syntactic structure to indicate whether a container is used to mean a measure of unit of some substance, in which the the noun for the container and for the substance are said consecutively. Thus, "et glas vand" (glossed as "a glass water") means "a glass *of* water". This structure, represented in the Danish treebank as the content being a dependent of the container, is also a cause of systematic disagreement.

(9.3)
1. In a second operation, mixing tools used to make urethane are soaked or agitated **in** *buckets* of dibasic esters (instead of methylene chloride) [...]
2. El informe relata que la embarcación también transportaba **medio** *saco* de carbón [...]
   (The report tells the ship also transported **half** a *sack* of coal [...])
3. Han forestiller sig, at *æsk**en*** har vinger.
   (He imagines **the**-*box* has wings.)
4. Inden jeg langt om længe rejser mig op for at øve mig i at gå med en stor *skål* danske **jordbær** oven på hovedet.
   (Before I finally stand up to practice walking with a large *bowl* (**of**) Danish **strawberries** on the-head.)

9.4 Conclusions

In this chapter we have described a system to predict the observed agreement A_o and automatically determine examples that have low agreement. The lower R^2 scores for agreement prediction than for for literality prediction determine that this task is more difficult.

Nevertheless, we have been able to fit six datasets over baseline, namely the three language variants of the CONTCONT and LOCORG datasets. The datasets that cannot be fit are the two lowest-agreement datasets, and the highest-agreement dataset, which has too little variation in A_o for the relation between dependent variable and features to be modelable.

Most syntactic features correlate with high agreement, which implies that marked syntactic contexts help an univocal interpretation. Most negative features for agreement are lexical, which indicates that difficult or unusual words get on the way of annotator agreement.

Nevertheless, we have found linguistic cues that pinpoint to syntactic behavior of the headword that cause agreement to drop, because annotators interpret them in different manners.

There are no features for the headword, because we are using features from the class-based WSD and the headword is abstracted away. Including the headword would improve R^2 because a great deal of the variation in agreement between datasets is caused by the headword and the semantic class, regardless of context.

This system can be used as an automatic review method of for sense-annotation tasks, because it identifies a proportion of systematic disagreement that can be attributed to certain linguistic features, which can lead to reviews of annotation guidelines.

Chapter 10

Conclusions

This chapter introduces in Section 10.1 the conclusions to the research question stated in Section 2.7, plus it lists the contributions of this thesis beyond answering the question about the adequacy of the token-level representation of the underspecified sense in Section 10.2. Lastly, Section 10.3 outlines the future work lines that can be followed to complement the results of this dissertation.

10.1 Main conclusions

In this section we take up the main research question in this dissertation, namely whether there is enough empirical evidence for the inclusion of an underspecified sense in our sense inventories dealing with dot-type nominals.

We conclude that a discrete, independent representation for the underspecified sense is not desirable for the representation of the senses of dot-type words at the token level. In the rest of this section we break down the argument with regards to the specific questions defined in Section 2.7.

Human judgments

In Chapter 4 we have analyzed the human judgments on sense underspecification. Comparing the behavior between turkers and volunteers it can determined that the underspecified sense cannot be explicitly captured by turkers using the annotation procedure, but that there is significant ($> 5\%$) amount of simple-majority underspecified senses for the volunteer-annotated datasets.

This indicates that the sense inventory is task-setup dependent, and we cannot identify the underspecified sense explicitly from turkers.

Word-sense induction

In Chapter 5 we have analyzed the distributional behavior of dot type words using word sense induction (WSI). The WSI system has not been able to find distributional evidence for the underspecified sense in our WSI experiments. Nevertheless, WSI has neither been able to identify the literal or metonymic sense, except those few examples that are marked by decisive features like certain prepositions.

We could dismiss the underspecified sense as a distributional phenomenon if the system had properly identified the literal and metonymic senses, thus knowing our system was adequate for the task. Therefore, we consider our WSI experiment non-conclusive on whether there is sufficient distributional evidence to postulate the underspecified sense for token-sense representation.

However, given the results in word-sense disambiguation (see below), we predict that the underspecified sense would not be identifiable using WSI, regardless of the success at identifying the literal and metonymic senses.

Word-sense disambiguation

In Chapter 7 we have analyzed the performance of a classification system to predict the literal, metonymic and underspecified senses. This system is an instance of a figurative-language resolution system, which is a kind of word-sense disambiguation (WSD).

Our WSD system has been able to predict the alternating literal and metonymic senses with acceptable performance (F1 scores as high as 0.88 for literal and 0.77 for metonymic), but the performance for the underspecified sense was very low (F1 scores of at most 0.17). Thus, the WSD experiments do not justify postulating an independent, underspecified sense.

Alternative representation for dot-type sense

We have introduced an alternative representation for the senses in a dot type as a continuous value between 0 (fully metonymic) to 1 (fully literal).

We successfully fit a regression model for each dataset using the literality score as a dependent variable. In a comparative study with WSD, we deem the literality score to be a more robust representation, and thus we consider this literality score LS an adequate candidate for the representation.

10.2 Contributions

The main contribution of this dissertation is an empirical study on the human ability to recognize underspecified senses for regular polysemy, and on the ability of NLP system to reproduce these human judgments. We have conducted the study for five dot types in English, and chosen two relevant dot types (CONTAINER•CONTENT and LOCATION•ORGANIZATION) for a cross-linguistic study that also includes Danish in Spanish.

This dissertation offers an expansion on four articles that relate to the topic of sense underspecification in regular polysemy. In Chapter 4 we expand on Martínez Alonso et al. (2013), focusing on the method to obtain human judgments on regular polysemy, describing the SAMs more extensively, providing examples and settling for a SAM to conduct the other experiments. The data described in Martínez Alonso et al. (2013) is freely available on MetaShare[1].

In Chapter 5 we expand on Romeo et al. (2013) by running WSI on all datasets, and not only on English. In Chapter 7 we also expand the work in Martínez Alonso et al. (2011) and Martínez Alonso et al. (2012) by running WSD on all datasets, as well as improving the feature space.

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

We have proposed the literality score (LS), a continuous representation for the senses where 0 is fully metonymic and 1 is fully literal, instead of a representation where literal, metonymic and underspecified are discrete, disjoint categories. When used as a continuous dependent variable to replace the sense tags, LS prediction is a more robust method to characterize the token-wise sense behavior of dot-type words, provided that these words belong to dot types where there are enough intermediate readings—unlike, for instance, ANIMAL•MEAT.

Postulating that the literality coefficient is a more adequate representation for the token-wise sense behavior of dot-type words does not necessarily violate the representation of dot types as lexical entries, even though it offers new possibilities for the annotation of these words which are not compliant with the GLML (cf. Section 3.2.1).

We consider that the robustness of the literality coefficient is supporting argument against the sense-enumeration paradigm, as we have seen that explicitly representing the underspecified sense does not necessarily aid the performance of our WSD systems.

In the analysis of human annotations, we have conducted a study on the biases of the different kinds of available annotators (expert, turkers, volunteers). We have established that turkers disprefer the underspecified sense, and that the expert can be overzealous when identifying it.

With regards to the evaluation of our metonymy-resolution WSD system, we have conducted a critical analysis of the baselines used in the SemEval2007 shared task and determined that they do not lend themselves well to being used as baselines for any metonymic sense alternation because they are tailored for the metonymies derived from LOCATION words.

Moreover, this dissertation also has used approaches that are, to the best of our knowledge, the first attempts at conducting certain tasks. Our WSI method, which has not yielded very good results, is a first attempt at implementing class-based WSI for dot-type words.

Likewise, we have presented an experiment to determine the agreement of our sense-annotated examples, in order to capture which proportion of the disagreement is systematic—i.e. correlates with the linguistic features of the example—and can be predicted.

10.3 Further work

The results of this dissertation set new lines of research, as well as suggesting methodological improvements on some parts of our work. In this section we outline possibilities for future work that builds on the results of this dissertation.

With regards to **the human recognition** of the underspecified sense, we want to obtain volunteer annotations for English. This would allow comparing the difference in distributions with regards to turkers and experts. Ideally, the volunteers for English would yield sense distributions that resemble those obtained from the volunteers for Danish and Spanish.

If the volunteer annotations are indeed the most representative ones, a linguistically informed method could be developed that predicted the volunteer sense assignments from an input composed of linguistic features and the judgments of turkers.

We have used a simple, threefold sense inventory for our annotation task for both turkers and volunteers, even though the expert annotation scheme is closer to a two-sense annotation scheme where multiple tagging is possible: if the paraphrases for both literal and metonymic were either possible or necessary, then the underspecified sense was assigned. Our annotation scheme for turkers and volunteers could be contrasted with a two-sense scheme with multiple tagging, although this would come with a revision of our sense-assignment method.

We also consider the possibility of developing a sense-assignment method that relies both on the theoretical assumption behind VOTE and the latent-variable approach used by MACE. A first approach would be a method using Bayesian inference instead of EM—used by MACE—, because Bayesian methods yield more skewed distributions that are closest to the sense distributions we expect.

With regards to the **WSI experiments**, our ability to draw conclusions was limited by the performance of the system. It is be desirable to repeat the experiments using a richer feature space that uses parsed corpora as input and also incorporates inflectional morphology.

Finally, we want to assess the **usefulness of the literality score** as a replacement for the three-way representation of dot-type senses. Using extrinsic evaluation, we can use dot-sense annotated input for an NLP task like machine translation, question answering or information retrieval, and measure whether the system fares best with a discrete or continuous representation.

However, as we show in Elming et al. (2013), when conducting extrinsic evaluation of dependency-tree formalisms, the external NLP tasks also have biases of their own, and it would be necessary to experiment with several tasks to get a fair overview of which representation—continuous or discrete—yields systematically better results.

Bibliography

Steven Abney. *Semisupervised learning for computational linguistics*. CRC Press, 2010.

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June 2007.

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

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

Béatrice Arnulphy, Xavier Tannier, and Anne Vilnat. Event nominals: Annotation guidelines and a manually annotated corpus in french. In *LREC*, pages 1505–1510, 2012.

Javier Artiles, Julio Gonzalo, and Satoshi Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

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

N. Asher and J. Pustejovsky. Word meaning and commonsense metaphysics. *Semantics Archive*, 2005.

Nicholas Asher. *Lexical meaning in context: A web of words*. Cambridge University Press, 2011.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009.

Paz Battaner. El fenómeno de la polisemia en la lexicografía actual: otra pespectiva. *Revista de lexicografía*, 14:7–25, 2008.

Núria Bel, Maria Coll, and Gabriela Resnik. Automatic detection of non-deverbal event nouns for quick lexicon production. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 46–52. Association for Computational Linguistics, 2010.

Núria Bel, Lauren Romeo, and Muntsa Padró. Automatic lexical semantic classification of nouns. *arXiv preprint arXiv:1303.1930*, 2013.

Melanie J. Bell and Martin Schäfer. Semantic transparency: challenges for distributional semantics. In *Proceedings of IWCS 2013 Workshop Towards a Formal Distributional Semantics*, pages 1–10, Potsdam, Germany, March 2013. Association for Computational Linguistics.

Chris Biemann and Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28. Association for Computational Linguistics, 2011.

Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*, 2006.

Reinhard Blutner. Lexical pragmatics. *Journal of Semantics*, 15(2):115–162, 1998.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. Modelling polysemy in adjective classes by multi-label classification. In *EMNLP-CoNLL*, pages 171–180, 2007.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. An analysis of human judgements on semantic classification of catalan adjectives. *Research on Language and Computation*, 6(3-4):247–271, 2008.

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

Gemma Boleda, Padó Sebastian, and Jason Utt. Regular polysemy: A distributional model. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 151–160, Montréal, Canada, 7-8 June 2012b. Association for Computational Linguistics.

Jacques Bouaud, Bruno Bachimont, and Pierre Zweigenbaum. Processing metonymy: a domain-model heuristic graph traversal approach. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 137–142. Association for Computational Linguistics, 1996.

P. Bouillon, V. Claveau, C. Fabre, and P. Sébillot. Acquisition of qualia elements from corpora-evaluation of a symbolic learning method. In *3rd International Conference on Language Resources and Evaluation, LREC*, volume 2. Citeseer, 2002.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

Paul Buitelaar. *CoreLex: systematic polysemy and underspecification.* PhD thesis, Citeseer, 1998.

Chris Callison-Burch. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics, 2009.

José M. Castaño. *Spanish Clitics, Events and Opposition Structure*, pages 147–179. Advances in Generative Lexicon Theory. Springer, 2013.

Nancy Chinchor and Patricia Robinson. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, 1997.

Massimiliano Ciaramita and Yasemin Altun. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602. Association for Computational Linguistics, 2006.

Stephen Clark and David Weir. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

Vincent Claveau and Pascale Sébillot. Automatic acquisition of gl resources, using an explanatory, symbolic technique. In *Advances in Generative Lexicon Theory*, pages 431–454. Springer, 2013.

A. Copestake. Can distributional approached improve on goold old-fashioned lexical semantics? In *IWCS Workshop Towards a Formal Distributional Semantics.*, 2013.

Ann Copestake and Ted Briscoe. Semi-productive polysemy and sense extension. *Journal of semantics*, 12(1):15–67, 1995.

William Croft. The role of domains in the interpretation of metaphors and metonymies. *Cognitive linguistics*, 4(4):335–370, 1993.

D. A. Cruse. *Lexical semantics*. Cambridge Univ Pr, 1986.

James R Curran. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 26–33. Association for Computational Linguistics, 2005.

Gerard de Melo, Collin F Baker, Nancy Ide, Rebecca J Passonneau, and Christiane Fellbaum. Empirical comparisons of masc word sense annotations. In *LREC*, pages 3036–3043, 2012.

Anca Dinu. Building a generative lexicon for romanian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta*, 2010.

René Dirven. Metonymy and metaphor: Different mental strategies of conceptualization. *Metaphor and metonymy in comparison and contrast*, 112, 2002.

William B Dolan. Metaphor as an emergent property of machine-readable dictionaries. *Proceedings of Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity*, pages 27–29, 1995.

Markus Egg, Joachim Niehren, Peter Ruhrberg, and Feiyu Xu. Constraints over lambda-structures in semantic underspecification. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 353–359. Association for Computational Linguistics, 1998.

Jakob Elming, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Hector Martinez, and Anders Søgaard. Down-stream effects of tree-to-dependency conversions. In *Proceedings of NAACL-HLT*, pages 617–626, 2013.

Ulle Endriss and Raquel Fernández. Collective annotation of linguistic resources: Basic principles and a formal model. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 539–549, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Maria Victoria Escandell-Vidal. *Fundamentos de semántica composicional*. Ariel, 2004.

Dan Fass. Metonymy and metaphor: what's the difference? In *Proceedings of the 12th conference on Computational linguistics-Volume 1*, pages 177–181. Association for Computational Linguistics, 1988.

Dan Fass. met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90, 1991.

C. Fellbaum. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

W. A. Gale, K. W. Church, and D. Yarowsky. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, page 237. Association for Computational Linguistics, 1992.

Raymond W Gibbs. Literal meaning and psychological theory. *Cognitive science*, 8(3):275–304, 1984.

Daniele Godard and Jacques Jayez. Towards a proper treatment of coercion phenomena. In *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics*, pages 168–177. Association for Computational Linguistics, 1993.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. Multilingual central repository version 3.0. In *LREC*, pages 2525–2529, 2012.

H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

J. B. Grimshaw. *Argument structure*. MIT press Cambridge, MA, 1991.

Patrick Hanks. The syntagmatics of metaphor and idiom. *International Journal of Lexicography*, 17(3):245–274, 2004.

Patrick Hanks. Metaphoricity is gradable. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 171:17, 2006.

Sanda Harabagiu. Deriving metonymic coercions from wordnet. In *Workshop of the Usage of WordNet in Natural Language Processing Systems, COLING-ACL*, volume 98, pages 142–148, 1998.

Jerry Hobbs. Syntax and metonymy. *The Language of Word Meaning*, page 290, 2001.

Jerry R. Hobbs and Paul Martin. Local pragmatics. *Proceedings, International Joint Conference on Artificial Intelligence*, 1987.

Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63(1):69–142, 1993.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with mace. In *Proceedings of NAACL-HLT*, pages 1120–1130, 2013.

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

Seohyun Im and Chungmin Lee. *Combination of the Verb Ha-'Do'and Entity Type Nouns in Korean: A Generative Lexicon Approach*, pages 203–226. Advances in Generative Lexicon Theory. Springer, 2013.

Panagiotis Ipeirotis. Demographics of mechanical turk. In *NYU working paper no*, 2010.

Ann Irvine and Alexandre Klementiev. Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 108–113, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Rubén Izquierdo, Armando Suárez, and German Rigau. An empirical study on class-based word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 389–397. Association for Computational Linguistics, 2009.

Roman Jakobson. The metaphoric and metonymic poles. *Fundamenals of language*, 1971.

Per Anker Jensen and Carl Vikner. Leksikalsk semantik og omverdensviden. In *Sprogteknologi i Dansk Perspektiv - En samling artikler om sprogforskning og automatisk sprogbehandling*. C.A. Reitzels Forlag, 2006.

Elisabetta Ježek and Alessandro Lenci. When gl meets the corpus: a data-driven investigation of semantic types and coercion phenomena. *Proceedings of GL*, pages 10–11, 2007.

Elisabetta Ježek and Chiara Melloni. Nominals, polysemy, and co-predication*. *Journal of Cognitive Science*, 12:1–31, 2011.

Elisabetta Ježek and Valeria Quochi. Capturing coercions in texts: a first annotation exercise. In *LREC*, 2010.

Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367, 2008.

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

Anders Johannsen, Hector Martinez Alonso, Christian Rishøj, and Anders Søgaard. Shared task system description: Frustratingly hard compositionality prediction. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 29–32. Association for Computational Linguistics, 2011.

Michael Johnston and Federica Busa. Qualia structure and the compositional interpretation of compounds. In *Proceedings of the ACL SIGLEX workshop on breadth and depth of semantic lexicons*, pages 77–88, 1996.

D. Jurgens and K. Stevens. Measuring the impact of sense similarity on word sense induction. In *First workshop on Unsupervised Learning in NLP (EMNLP 2011)*, 2010.

David Jurgens. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 556–562, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

Adam Kilgariff and Martha Palmer, editors. *Proceedings of the Pilot SensEval*. Association for Computational Linguistics, Hermonceux Castle, Sussex, UK, September 1998. URL `http://www.aclweb.org/anthology/S98-1`.

A. Kilgarriff. " i don't believe in word senses". *Computers and the Humanities*, 31(2):91–113, 1997.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. Itri-04-08 the sketch engine. *Information Technology*, 105:116, 2004.

W. Kintsch. Predication. *Cognitive Science*, 25(2):173–202, 2001.

Walter Kintsch. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2):257–266, 2000.

Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Klaus Krippendorff. Agreement and information in the reliability of coding. *Communication Methods and Measures*, 5(2):93–112, 2011.

Matthias Trautner Kromann. The danish dependency treebank and the dtag treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT)*, page 217, 2003.

G. Lakoff and M. Johnson. *Metaphors we live by*. Chicago London, 1980.

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

Mirella Lapata and Chris Brew. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73, 2004.

Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, et al. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.

Johannes Leveling and Sven Hartrumpf. On metonymy recognition for geographic information retrieval. *International Journal of Geographical Information Science*, 22(3):289–299, 2008.

Jianguo Li and Chris Brew. Class-based approach to disambiguating levin verbs. *Natural Language Engineering*, 16(4):391–415, 2010.

Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics, 2010.

Percy Liang. Semi-supervised learning for natural language. Master's thesis, Massachusetts Institute of Technology, 2005.

Birte Loenneker-Rodman and Srini Narayanan. Computational approaches to figurative language. *Cambridge Encyclopedia of Psycholinguistics*, 2010.

John Lyons. *Semantics. vol. 2*. Cambridge University Press, 1977.

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Montserrat Marimon, Beatriz Fisas, Núria Bel, Marta Villegas, Jorge Vivaldi, Sergi Torner, Mercè Lorente, Silvia Vázquez, and Marta Villegas. The iula treebank. In *LREC*, pages 1920–1926, 2012.

Katja Markert and Udo Hahn. Understanding metonymies in discourse. *Artificial Intelligence*, 135(1):145–198, 2002.

Katja Markert and Malvina Nissim. Metonymy resolution as a classification task. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 204–213. Association for Computational Linguistics, 2002a.

Katja Markert and Malvina Nissim. Towards a corpus annotated for metonymies: the case of location names. In *LREC*. Citeseer, 2002b.

Katja Markert and Malvina Nissim. Metonymic proper names: A corpus-based account. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 171: 152, 2006.

Katja Markert and Malvina Nissim. Semeval-2007 task 08: metonymy resolution at semeval-2007. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 36–41, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

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

Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC Press, 2011.

Ricardo Martin-Brualla, Enrique Alfonseca, Marius Pasca, Keith Hall, Enrique Robledo-Arnuncio, and Massimiliano Ciaramita. Instance sense induction from attribute sets. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 819–827, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1944566.1944660`.

David Martínez, Eneko Agirre, and Lluís Màrquez. Syntactic features for high precision word sense disambiguation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

Héctor Martínez Alonso, Núria Bel, and Bolette Sandford Pedersen. A voting scheme to detect semantic underspecification. In *LREC*, pages 569–575, 2012.

Héctor Martínez Alonso, Bolette Sandford Pedersen, and Núria Bel. Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–730, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

Héctor Martínez Alonso, Núria Bel, and Bolette Sandford Pedersen. Identification of sense selection in regular polysemy using shallow features. In *Proceedings of the 18th Nordic Conference of Computational Linguistics, Riga (NoDaLiDa 2011)*, NEALT Proceedings Series, 2011.

Zachary J Mason. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44, 2004.

Andrew Kachites McCallum. *Mallet: A machine learning for language toolkit*. http://mallet.cs.umass.edu, 2002.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics, 2004.

B. Mellebeek, F. Benavent, J. Grivolla, J. Codina, M. R. Costa-Jussa, and R. Banchs. Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 114–121. Association for Computational Linguistics, 2010.

Paola Merlo and Suzanne Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*, 27 (3):373–408, 2001.

Rada Mihalcea. Co-training and self-training for word sense disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2004)*, 2004.

Masaki Murata, Qing Ma, Atsumu Yamamoto, and Hitoshi Isahara. Metonymy interpretation using x no y examples. *arXiv preprint cs/0008030*, 2000.

Michal Měchura. *Selectional Preferences, Corpora and Ontologies*. PhD thesis, Ph. D. thesis, Trinity College, University of Dublin, 2008.

Vivi Nastase, Alex Judea, Katja Markert, and Michael Strube. 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, 2012.

Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

Roberto Navigli and Daniele Vannella. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, USA, June 2013a. Association for Computational Linguistics.

Roberto Navigli and Daniele Vannella. Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, Georgia, USA, June 2013b. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

S. Nimb and B. S. Pedersen. Treating metaphoric senses in a danish computational lexicon: Different cases of regular polysemy. *Ulrich Heid, Stefan Evert, Egbert Lehmann, Christian Rohrer (eds.)*, pages 679–692, 2000.

Malvina Nissim and Katja Markert. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 56–63. Association for Computational Linguistics, 2003.

Malvina Nissim and Katja Markert. Annotation scheme for metonymies. In *Technical document*, 2005a.

Malvina Nissim and Katja Markert. 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*, 2005b.

Pierre Nugues and Kalep Heiki-Jaan. Extended constituent-to-dependency conversion for english. *NODALIDA 2007 Proceedings*, pages 105–112, 2007.

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

Naoyuki Ono. *On the Event Structure of Indirect Passive in Japanese*, pages 311–326. Advances in Generative Lexicon Theory. Springer, 2013.

Boyan Onyshkevych. Nominal metonymy processing. *Atelier Coling–Acl*, 98, 1998.

Nicholas Ostler and Beryl T Sue Atkins. Predictable meaning shift: Some linguistic properties of lexical implication rules. In *Lexical Semantics and knowledge representation*, pages 87–100. Springer, 1992.

Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.

Rebecca J. Passonneau and Bob Carpenter. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. The masc word sense sentence corpus. In *Proceedings of LREC*, 2012.

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

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Lars Trap-Jensen, and Henrik Lorentzen. Dannet-a wordnet for danish. In *Proceedings of the Third International WordNet Conference*, 2006.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

Wim Peters. Metonymy as a cross-lingual phenomenon. In *Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14*, pages 1–9. Association for Computational Linguistics, 2003.

Winn Peters and Ivonne Peters. Lexicalised systematic polysemy in wordnet. In *Proc. Secondt Intnl Conf on Language Resources and Evaluation*, 2000.

Manfred Pinkal and Michael Kohlhase. Feature logic for dotted types: A formalism for complex word meanings. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 521–528. Association for Computational Linguistics, 2000.

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

Group Pragglejaz. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39, 2007.

Judita Preiss and David Yarowsky, editors. *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, Toulouse, France, July 2001.

J. Pustejovsky. *The generative lexicon: a theory of computational lexical semantics*. MIT Press, 1995.

J. Pustejovsky. *Type theory and lexical decomposition*, pages 9–38. Springer, 2006.

J. Pustejovsky and P. G. Anick. On the semantic interpretation of nominals. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 518–523. Association for Computational Linguistics Morristown, NJ, USA, 1988.

J. Pustejovsky and B. Boguraev. *Lexical semantics: The problem of polysemy*. Clarendon Press, 1996.

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

James Pustejovsky. Type construction and the logic of concepts. *The language of word meaning*, 91123, 2001.

James Pustejovsky and Elisabetta Ježek. Semantic coercion in language: Beyond distributional analysis. *Italian Journal of Linguistics*, 20(1):175–208, 2008.

James Pustejovsky, Anna Rumshisky, Jessica L Moszkowicz, and Olga Batiukova. Glml: Annotating argument selection and coercion. In *Proceedings of the Eighth International Conference on Computational Semantics*, pages 169–180. Association for Computational Linguistics, 2009b.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010.

François Recanati. The alleged priority of literal interpretation. *Cognitive science*, 19(2):207–232, 1995.

François Recanati. Literal/nonliteral. *Midwest Studies in Philosophy*, 25:264–274, 2002.

Marta Recasens and Marta Vila. On paraphrase and coreference. *Computational Linguistics*, 36(4):639–647, 2010.

Marta Recasens, Maria Antònia Martí, and Constantin Orasan. Annotating near-identity from coreference disagreements. In *LREC*, pages 165–172, 2012.

Kirk Roberts and Sanda Harabagiu. Utdmet: Combining wordnet and corpus data for argument coercion detection. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 252–255, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

D. Rohde, L. Gonnerman, and D. Plaut. An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*, 2009.

Lauren Romeo, Héctor Martínez Alonso, and Núria Bel. Class-based word sense induction for dot-type nominals. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon*, 2013.

A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2007)*, 2007.

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

Anna Rumshisky, J Moszkowicz, and M Verhagen. The holy grail of sense definition: Creating a sense disambiguated corpus from scratch. In *Proceedings of 5th international conference on generative approaches to the lexicon (gl2009)*, 2009.

Roy Schwartz, Omri Abend, and Ari Rappoport. Learnability-based syntactic annotation design. In *COLING*, pages 2405–2422, 2012.

Ekaterina Shutova. 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, 2009.

Ekaterina Shutova and Simone Teufel. Metaphor corpus annotated for source-target domain mappings. In *LREC*, 2010.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 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, 2008.

Caroline Sporleder and Linlin Li. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics, 2009.

David Stallard. Two kinds of metonymy. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 87–94. Association for Computational Linguistics, 1993.

Jiping Sun, Fakhri Karray, Otman Basir, and Mohamed Kamel. Natural language understanding through fuzzy logic inference and its application to speech recognition. In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, volume 2, pages 1120–1125. IEEE, 2002.

Jiping Sun, Khaled Shaban, Sushil Podder, Fakhri Karry, Otman Basir, and Mohamed Kamel. Fuzzy semantic measurement for synonymy and its application in an automatic question-answering system. In *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*, pages 263–268. IEEE, 2003.

Eve Sweetser. From etymology to pragmatics: The mind-body metaphor in semantic structure and semantic change. *Cambridge: CUP*, 1990.

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

Noriko Tomuro. Systematic polysemy and interannotator disagreement: Empirical examinations. In *Proceedings of the First International Workshop on Generative Approaches to Lexicon*, 2001a.

Noriko Tomuro. Tree-cut and a lexicon based on systematic polysemy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001b.

P.D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690, 2011.

Jason Utt and Sebastian Padó. Ontology-based distinction between polysemy and homonymy. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 265–274. Citeseer, 2011.

Jason Utt, Alessandro Lenci, Sebastian Padó, and Alessandra Zarcone. The curious case of metonymic verbs: A distributional characterization. In *Proceedings of the IWCS 2013 Workshop Towards a Formal Distributional Semantics*, 2013.

Cornelia Verspoor. Conventionality-governed logical metonymy. In *Proc. of the 2nd International Workshop on Computational Semantics*, pages 300–312. Citeseer, 1997.

Marta Vila Rigat, Maria Antònia Martí Antonín, and Horacio Rodríguez Hontoria. Paraphrase concept and typology: a linguistically based and computationally oriented approach. In *Procesamiento del lenguaje natural 46*. Sociedad Española para el Procesamiento del Lenguaje Natural, 2011.

Jorge Vivaldi. 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, 2009.

Piek Vossen. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston, 1998.

Piek Vossen, Wim Peters, and Julio Gonzalo. Towards a universal index of meaning. *Proceedings of SIGLEX99: Standardizing Lexical Resources*, 1999.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, pages 1–23, 2013.

Beatrice Warren. An alternative account of the interpretation of referential metonymy and metaphor. *Metaphor and metonymy in comparison and contrast*, pages 113–130, 2002.

Ichiro Yamada, Timothy Baldwin, Hideki Sumiyoshi, Masahiro Shibata, and YAGI Nobuyuki. Automatic acquisition of qualia structure from corpus data. *IEICE transactions on information and systems*, 90(10):1534–1541, 2007.

D. Yarowsky. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, page 271. Association for Computational Linguistics, 1993.

Alessandra Zarcone and Sebastian Padó. I like work: I can sit and look at it for hours. In *Type clash vs. plausibility in covert event recovery. In Proc. VERB 2010 workshop. Pisa, Italy*, 2010.

Alessandra Zarcone, Jason Utt, and Alessandro Lenci. Logical metonymy from type clash to thematic fit. In *AMLaP 2012 Conference*, page 18, 2012.

Appendices

Appendix A

Appendices to annotation scheme

A.1 Words for English

The following words have been either obtained from Rumshisky et al. (2007) or from using WordNet as a thesaurus to find synonyms for the words in Rumshisky et al.'s work.

Dataset	words
ENG:ANIMEAT	anchovy, bison, boar, buffalo, bull, calf, camel, carp, catfish, cattle, chicken, clam, cod, cow, crab, crayfish, deer, dog, donkey, duck, eel, elk, goat, goose, hare, hen, herring, hog, lamb, lobster, mackerel, mussel, octopus, ostrich, oyster, pheasant, pig, pigeon, prawn, quail, rabbit, reindeer, rooster, salmon, sardine, shrimp, snail, snake, squid, swine, trout, tuna, turkey, whale, yak
ENG:ARTINFO	book, CD, diary, dictionary, letter
ENG:CONTCONT	bag, bottle, bucket, container, crate, dish, flask, glass, jar, keg, kettle, pint, plate, pot, spoon, vessel, vial
ENG:LOCORG	Afghanistan, Africa, America, Boston, California, Canada, China, England, Europe, Germany, London
ENG:PROCRES	acquisition,approval, classification, construction, damage, discount, illustration, imitation, instruction, invention, purchase, reference, simulation

A.2 Words for Danish and Spanish

Danish or Spanish words that share an English translation—e.g. both *æske* and *skrin* translate as *box*— have a subindex in their English translation. The middle words are marked in bold to ease checkup.

Dataset	words	translation
DA:CONTCONT	æske, beholder, container, dåse, flaske, fustage, glas, gryde, **kande, kar, kasse, kop, krus, kurv, pose, sæk**, skål, skrin, tønde, trug	box$_1$, container$_1$, container$_2$, can, bottle, keg, glass, pot, **jug, box$_2$, cup, mug, basket, bag, sack**, bowl, box$_3$, barrel, trug
DA:LOCORG	Afghanistan, Afrika, Amerika, Århus, Berlin, Boston, **Danmark, England, Europa, Frankrig, Fyn, Japan, Jylland, Kina**, København, London, Paris, Tyskland	Afghanistan, Africa, America,Aarhus, Berlin, Boston, **Denmark, England, Europe, France, Funen, Japan, Jutland, China**, Copenhagen, London, Paris, Germany
SPA:CONTCONT	barrica, barril, bol, bolsa, bote, botella, botellín, **caja, cajetilla, cesta, cesto, contenedor, cubo, envase**, frasco, lata, olla, paquete, saco, taza, tazón	wine cask, barrel, bowl, bag, jar, bottle, phial, **box, package$_1$, basket$_1$, basket$_2$, container, bucket, tin**, flask, can, pot, package$_2$, sack, cup, mug
SPA:LOCORG	Afganistán, África, Alemania, América, Argentina, Barcelona, **California, Cataluña, China, Colombia, España, Europa, Francia, Inglaterra**, Japón, Londres, Madrid, México, París, Valencia	Afghanistan, Africa, Germany, America, Argentina, Barcelona, **California, Catalonia, China, Colombia, Spain, Europa, France, England**, Japan, London, Madrid, Mexico, Paris, Valencia

A.3 Synthetic examples to assess validity of AMT

This appendix lists the five synthetic examples used as a preliminary study to assess the adequacy of using Amazon Mechanical Turk to annotate the English data. This section provides the examples and breaks down the amount of annotations for each sense, the resulting A_o and the gold standard value in Table A.1.

(A.1) a) I killed and cooked the *kangaroo* myself.
b) Our *donkeys* were always free to pasture.
c) The soup had chunks of *turtle* floating in the broth.
d) The *turtle* was slow.
e) The *donkey* was delicious.

(A.2) a) I am reading a good *book* on economics.
b) He used a large *book* as a doorstop.
c) They had transferred all the *recordings* to a different shelf.
d) We found a stash of really interesting *magazines* from the 70s.

(A.3) a) Sarah kept out of sight, then followed at a careful distance, swinging a *basket*.
b) He spilled his *cup* of coffee.
c) We had to carry the *box* once it was full.
d) The *box* was remarkably large.
e) John ate the whole *can*.

(A.4) a) The weather in *Spain* is generally very hot in the summer.
b) *Japan* takes a conservative turn to its tax policies.
c) We are several time zones away from *Australia* .
d) The reform that *Italy* suggest is unacceptable.
e) This kind of landscape is typical from Mexico and other dry areas.

(A.5) a) The doctor's *examination* of the patient was successful.
b) The *exam* was long.
c) The *exam* was on the table.
d) The *development* was applauded.
e) The enemy's *destruction* of the city was awful to watch.

Example	L	M	U	A_0	GS
A.1 ENG:ANIMEAT					
a)	0	5	5	0.45	U
b)	10	0	0	1.0	L
c)	0	10	0	1.0	M
d)	10	0	0	1.0	L
f)	10	0	0	1.0	M
A.2 ENG:ARTINFO					
a)	3	7	0	0.54	M
b)	9	1	0	0.81	L
c)	5	2	3	0.34	M
d)	9	1	0	0.81	L
f)	3	5	2	0.34	U
A.3 ENG:CONTCONT					
a)	9	0	1	0.81	L
b)	3	4	3	0.22	M
c)	7	0	3	0.54	U
d)	8	0	2	0.71	L
f)	0	10	0	1.0	M
A.4 ENG:LOCORG					
a)	10	0	0	1.0	L
b)	2	8	0	0.71	M
c)	10	0	0	1.0	L
d)	0	10	0	1.0	M
f)	9	1	0	0.81	L
A.5 ENG:PROCRES					
a)	8	1	1	0.71	L
b)	8	0	2	0.71	L
c)	5	5	0	0.45	M
d)	2	5	3	0.35	M
f)	10	0	0	1.0	L

Table A.1: Evaluation of turker annotation for synthetic examples

A.4 Excerpts from annotated examples

This section provides the first 20 examples for each dataset, with array of annotation obtained from either turkers or volunters, and the sense tag assigned by VOTE or MACE.

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
tblqmssey	0	anchovy	There are still a few bona fide fishermen about,and if you can manage to drag yourself out of bed between six and eight in the morning,you ll see them returning to shore in their gaily painted,flat-bottomed wooden boats,with nets of sardines and	anchovies	0	U M L L L	L	L
tebzcagxs	1	bison	It has been argued that modern	bison	are descended from Beringian bison,but Cooper 's data suggest otherwise .	L L L L L	L	L
dihbgfmkq	2	boar	You may see a passing pleasure-boat on the water,or a jet-ski convoy on the lookout for wildlife such as herons,pelicans,deer,wild	boar	, even the occasional crocodile - though the local creatures tend to only be found in the most remote parts of the mangrove,well away from humans .	L L L L L	L	L
vldkpohxz	3	cod	In a region without any noteworthy industry other than	cod	fishing,a little forestry,and tourism,the people add to their income by selling their craftwork and farm produce from improvised stands .	L L L L M	L	L
rzuuwgrxj	4	crab	Kids love the Touch Tide Pool where they can handle	crabs	and starfish .	L L L L L	L	L
mhnhjbuxe	5	duck	Retracing your steps on the narrow street flanking the cathedral,circle around to the rear and	duck	into the narrow carrer del Parads .	L L L L L	L	L
pcbfetwuj	6	eel	If you d like to try fishing,you can get yourself a good barbecued supper of white fish,pike perch or	eel	from the Havel,Mggelsee,and the Glienicker See .	U M M M M	M	M
qpgxlegbl	7	elk	Researchers are working to determine the minimum incubation time of CWD before clinical signs appear,now roughly estimated at 15 months in deer and 12 - 34 months in	elk	0	L L L L L	L	L
ujnjjhzjz	8	goat	The park is one of the most visited parts of the Minho,yet it is very easy to hike up into the mountains and find yourself alone with the mountain	goats	0	L L L L L	L	L
tspnekzfy	9	hare	" He is referring to the fact that some teen-aged "" children "" were occasionally executed for crimes (the crime of murder in the period to which he refers,or perhaps for stealing"	hare	from the king 's wood in earlier times) .	L L M L L	L	L
cmxgabjou	10	hog	Landis Valley Museum will present volunteers for the Landis Valley Cookbook Committee,in traditional dress,demonstrating the various steps to preparing a	hog	for consumption .	U M M M U	M	M
poaxomjlw	11	lamb	While I 'd rather be discussing your superb piece in this morning 's Times on the limits of online grocery shopping (so that 's why we did n't have	lamb	shanks for dinner . . .)	M M M M M	M	M
eurwptaks	12	lobster	But if this is the general future of food,I 'll have the	lobster	0	M M M M M	M	M
waxlfswqh	13	mackerel	The town has a picturesque working harbor,where small,brightly painted fishing boats bob,and larger vessels haul in daily catches of lobster,eel,and	mackerel	0	L M U U U	U	U
nmuwgspbc	14	octopus	Although the Mediterranean Sea is rather overfished and the waters are not warm enough to support brightly colored tropical fish,there are still numerous varieties to spot in the rocky shallows,including	octopi	that make their homes in tiny crevices .	L L L L L	L	L
ndjnlpqhn	15	ostrich	" In South Africa,which recently banned Hong Kong chickens,the Johannesburg Star reported that a woman had been "" trampled and kicked "" to death by an"	ostrich	"at a farm near Cape Town while her husband,already badly injured by the same bird,lay helplessly by "" for hours "" on a dirt road . "	L L L L L	L	L
atovtotaw	16	oyster	(She does n't mention noticing any	oysters	.)	L L L L L	L	L
qjfahdqbr	17	pheasant	It might pass for deserted,if not for the peacocks and	pheasants	who inhabit the botanical gardens,and the visitors strolling amidst the palms and rhododendrons .	L L L L L	L	L
xhaiqikqx	18	pigeon	Old men sell bags of corn for you to feed to the	pigeons	, while brightly dressed water-sellers tout for business with cries of buz gibi !	L L L L L	L	L
rxvregfso	19	quail	" But when he ate well he ate it all up and said,"" My heavens,this ragout of"	quail	"is simply delicious . "" "	M M M M M	M	M

Table A.2: First twenty examples for ENG:ANIMEAT

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
cpqbcofjd	0	book	An article ridicules the recent spate of	books	on human behavior .	L L M U M	U	U
aajosilvt	151	diary	Up on the first floor is the Muse de l Histoire de France , with gems such as the only known portrait of Joan of Arc painted in her lifetime and the	diary	kept by Louis XVI .	L L M L L	L	L
isflfjrtd	243	dictionary	Although L has , in addition to the names , abbreviations , and only a few pages of miscellaneous materials- Handbook of Style , Ten Vexed points in English grammar-and some tables of moneys , weights and measures , etc . , W has short sections on Foreign Words and Phrases (most of which are dispensable) , Signs and Symbols , and Style , as well as one of those interminably boring listings of Colleges and Universities that clutter up most American college	dictionaries	0	M U M M L	M	M
ctjmpqoyz	247	dictionary	In short , just because a new word has been coined does not mandate its publication in	dictionaries	in the frequency to which we have become victim over the past thirty-odd years .	L U L M L	L	L
eapyqemvb	250	dictionary	Referring to	dictionaries	old and new , I found gedunk listed only in Webster 3 , with that frustrating label origin unknown .	L M M M M	M	M
cibdmoppe	159	diary	Athens comes alive after dark with a range of activities to fill your	diary	0	L L L L M	L	L
xbxsyetjz	253	dictionary	In the US , publishers do not generally advertise the number of definitions in their	dictionaries	but flaunt their entry counts .	M M U U L	M	U
rafttjajl	143	diary	If Husarska keeps this up , her	diary	might as well have been written from a shopping mall in Dayton .	L M M M M	M	M
pfjponnbr	139	diary	A key reason : The Fifth Amendment does not protect self-incriminating	diaries	because diary writing is voluntary , not compelled by the court .	M M U M U	M	U
rlbsgcbzz	147	diary	But then I read in Bob Haldeman 's	diary	, which was published in 1994 , that in 1972 Secretary of the Treasury John Connally had complained to Haldeman that several people–including me–were conspiring against him .	M U U M M	M	U
icqxnjuhk	125	diary	During sweep months , Nielsen sends out "	diaries	" to 250 , 000 households across all these markets .	L M L L M	L	L
ioezukmzt	121	diary	Second , speaking from experience of being at an Ivy and being in a fraternity , there are some things that are really missing in their	diaries	0	M L M M M	M	M
qyfetvjxj	134	diary	(Click here to read the	diary	of Pi star Sean Gullette .	M L M M M	M	M
dqaternmq	129	diary	Virtually any	diary	or utterance to a friend can now be used against you .	L L M M L	L	L
qrzysgvnd	292	book	There 's much in the elegiac tone of this	book	that reminds me of Philip Larkin 's collection All What Jazz ?	M M U M M	M	M
vugppzppy	277	book	Copies even found their way into the trenches of the First World War , so Frank Richards (his favorite name) stamped generations of boys from 1908 to 1940 and even later when the tales were turned into	books	, up until the last Bunter Story appeared in 1960 , shortly before Richards ' death .	L M L M M	M	L
dhmqqdokh	280	book	A delightful essay slams Germaine Greer 's new	book	as " a sour and undiscriminating litany of charges against men . "	M L L U M	U	U
asugsqqjt	283	book	The conclusion is that such	books	seldom have anything to teach us except in the most general way about the way the language was used in another time , partly owing to the lack of sophistication of their compilers , partly to their conservatism , which tempts them to include terms and definitions that are no longer current .	M M M M M	M	M
geybkafgm	287	book	It seemed best not to try to go through the	book	systematically but to dip into it here and there .	M M M L U	M	U
dsxrlqrdj	266	book	D . Barnes & Noble 's ownership of	book	distributor Ingram , says Avin Domnitz , head of the American Booksellers Association .	L L L M L	L	L

Table A.3: First twenty examples for ENG:ARTINFO

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
nmcqikvsh	494	container	However , these	containers	may be repeatedly reused for storing uncontaminated waters such as deionized or laboratory-prepared dilution waters and receiving waters .	L L L L L	L	L
dymmjgssq	470	bag	A Japanese carrier	bag	showing yachts on a blue sea had the message , Switzerland : Seaside City .	L L L L L	L	L
ohyuoctvg	391	flask	The remaining two	flasks	, designated infected (I) , were each inoculated with an aliquot of P . carinii - infected lung supernatant containing 7 10 5 P . carinii organisms per ml in 30 ml of media .	L M L U L	L	L
lbuvwakbs	404	flask	Two	flasks	, designated mock-infected (M) , were inoculated with an aliquot of normal rat lung homogenate in 30 ml of complete media .	L L U L U	L	U
dgcgmbkry	165	spoon	A German doctor named Lorenz Hoffman had a typical Kunst- und Wunderkammer : he owned paintings by Durer and Cranach , a skeleton of a newborn , two dozen miniature	spoons	hidden within a cherry pit , an armband made of elks ' hoofs , mummies and various rare musical instruments .	M L L L L	L	L
zrqxhzydc	190	spoon	She dropped her	spoon	into her coffee . "	L L L L L	L	L
jvisapjwh	265	plate	Aliquots of 20 l of each lysate were transferred to flat-bottomed , white microtiter	plates	(Dynex , Chantilly , VA , USA) and read on a Dynex MLX plate reader in flash mode .	M U L L L	L	L
xzuubvhcl	240	plate	The	plates	were washed thrice with PBS containing 0 . 05 % Tween 20 (Sigma Co .	L L L L L	L	L
lvkjolnvw	83	spoon	The Sun called him " a pampered , privileged young man who has achieved nothing , despite being born with a royal silver	spoon	in his mouth , despite all the undeserved advantages he has been handed . "	L L L L L	L	L
jqpholzpv	10	spoon	So , to hide them , I dropped a	spoon	on the floor .	L L L L L	L	L
pkuxuxkty	30	spoon	The maximum intake of non-breast milk fluid in the past week that was allowed in this definition of exclusive breastfeeding was set at 3 cucharillas (teaspoons) or 1 cuchara (larger	spoon	) or 1 / 2 fluid ounce .	L M L M M	M	M
azgtmbnin	21	spoon	(Structural linguists do not like to deal with meaning , either , satisfied that the meaning of a word is the sum of its contexts-a perfectly valid argument if one allows context to include one 's mother telling an infant , Here , darling , eat another	spoon	of applesauce .	M M M M L	M	M
ptsrupgcu	44	spoon	You ll also find lots of older pieces - not quite antique but still exquisite - in collectibles shops around the town in the form of pretty	spoons	, ornate pill and snuff boxes , or letter openers .	L L L L L	L	L
lgwhgxayd	36	spoon	That is , if you can stand the noise produced by wooden	spoons	hitting Tupperware .	L L L L L	L	L
pkobtudwi	63	spoon	Babies bang wooden	spoons	on overturned saucepans ; grannies rock so silently you can hear the creak of floorboards and hip bones .	L L L L L	L	L
cjnejjtpg	222	bucket	Let this groove lead to the bean field and guide the water spilled from the	bucket	such that it flows to water the bean field .	L M U M L	U	M
cljcsrmii	271	bag	Of course , George Bush was befuddled by a supermarket scanner , which is n't even a real hand tool , and besides , no one expected him to operate it , just carry the	bag	of groceries out to the car .	M U M L M	M	M
btwwuzjpw	276	bag	The only consequence of their selection was that their checked	bag	were held off the plane until it was confirmed that they had boarded the aircraft .	L L L U L	L	L
eobgsiqgl	286	dish	Cells were seeded at a density of 10 , 000 cells / cm 2 in polystyrene Falcon six-well plates or 60 mm	dishes	(Becton Dickinson Labware , Franklin Lakes , NJ) and maintained at 37 C under an atmosphere of 5 % of CO 2 and 95 % room air .	L M L L L	L	L
sxvrwinva	267	bag	The Japanese themselves (in the big cities at least) can be seen more often than not with some kind of shopping	bag	- they make very large , sturdy ones - just on the off chance that they might want to buy something .	L L L L L	L	L

Table A.4: First twenty examples for ENG:CONTCONT

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
tasnqmepy	0	Canada		Canada	s provincial tourist offices are conscientious and well worth consulting,even before you leave home .	M M M L U	M	M
fgxporyaq	289	America	But when you start counting by race , you divide	America	" , you do n't bring it together . """	M M M M M	M	M
ndzpltams	271	America	supporting	America	's publishing economy	U M M M M	M	M
bqbzvxlsm	272	America	In Nashville , he shows how country and western has become the predominant music of white	America	, with Garth Brooks having outsold every recording artist in the United States except the Beatles .	M L M M M	M	M
lirxrzpwm	287	America	An editorial darkly warns that China is modernizing and expanding its military in order to displace	America	as Asia 's dominant power .	M M M L M	M	M
krowhgssv	231	America	Mintz seems puzzled , musing that it was n't as though he 'd said	America	had no literature .	M M M M M	M	M
pklxlhhzb	235	America	China Daily said Friday in an editorial that behind	America	"'s "" barbaric air strikes against Yugoslavia "" lies a plan to divide Europe and dominate the world ."	M M M M M	M	M
wpbwbbmwd	246	America	Where King invoked	America	's promise , Malcolm underscored its betrayal .	U L M M M	M	M
wgpwpqaha	250	America	I have no idea about what he plans , but I have the awful feeling that he may have used me for a green card and for a way to keep his children in	America	0	L L L L L	L	L
snxusspgu	486	California	RNA was isolated using RNeasy Mini Kit (Qiagen , Valencia ,	California	, United States) , then subjected to DNase digestion followed by reverse transcription (Invitrogen) .	L L L L L	L	L
aayawdahf	494	California	Sections (5 m) mounted on Super Frost / plus glass (Menzel and Claser , both in Braunschweig , Germany) , were processed by a labelled streptavidin-biotin method with a Histostain Plus kit (Zymed , San Francisco ,	California	, USA) .	L L L L L	L	L
ptbuexuad	478	California	In the United States , at the National Institutes of Health in Bethesda , Maryland , the unaltered enzyme was infused directly into the venous circulation [2] ; at City of Hope in Duarte ,	California	, it was entrapped in red cell membranes coated with antibody in an effort to direct it to macrophages [3] .	L L L L L	L	L
xrkchheol	451	California	Held on a single day in the basement auditorium of the Berkeley Art Museum at the University of	California	at Berkeley , it brought together a typically motley collection of intellectuals who would willingly give over a sunny Saturday to an opportunity to learn from a panel of distinguished speakers .	L L L U M	L	L
lpllkzjtp	463	California	The nomination will be ratified at the August 2000 convention in	California	0	L L L L L	L	L
cdarpkmzj	497	England	It was James VI of Scotland who became James I of	England	in 1603 .	L U M M M	M	M
nkmzohbzp	260	Canada	If patriots make a reverent pilgrimage to	Canada	s Sports Hall of Fame (just off Lakeshore Boulevard West at Exhibition Place) to salute heroic athletes of the past , the separate Hockey Hall of Fame provides a tonic lesson for American visitors to recall , among the trophies , masks , skates , and hockey sticks , the fact that nearly all their ice hockey heroes are Canadian-born .	L L M M U	U	M
mgwukazio	278	Canada	Interestingly , the proportion of publications from Latin America , the United States , and	Canada	across all subject areas in Science and Nature were nearly identical to those of the top 20 ecological journals .	L L L U L	L	L
qqgzownkd	234	Canada	This takes some of the wind out of the sails of John Cabot , the Venetian navigator who claimed	Canada	s eastern seaboard for England s Henry VII in 1497 .	L U L L M	L	L
qwfvbpizh	482	England	Hawthorne gives Winslow an air of gorgeously ineffable sadness ; he seems to carry on his ever shakier shoulders the knowledge that this gray way of life , this	England	, is doomed .	M U M M M	M	M
avmvitfth	453	England	I never had any reason to suspect that the original was other then Darkies until I recently acquired a CD of songs from American musical shows which were recorded in	England	as performed there .	L L L L L	L	L

Table A.5: First twenty examples for ENG:LOCORG

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
kuehnryls	373	invention	Premiums could be given for therapies that address treatment gaps or for	inventions	that pave the way to new classes of drugs .	M M L M L	M	M
lquwfvwiu	297	simulation	The	simulation	is not a prediction of what will happen in the future .	L M M M M	M	M
nqybkxdji	320	simulation	It is important to point out that if a defibrillator is not readily available (the defibrillator was readily available in our	simulations	) , the time required to locate one and treat the patient could be much longer than that reported in our simulation .	L U L M M	U	L
jofxzgdfd	261	simulation	In fact , Loosemore 's computer	simulations	suggest that it will make flip-flops more common , not less .	L M U M L	U	L
ynmrpctqw	125	construction	For example , some of the clearest examples of niche	construction	occur in plant succession where , as F . E .	L L L M L	L	L
mocdmectm	441	construction	Torralba has perhaps the most beautiful taula of all - certain evidence that the ancients were masters of the art of stone	construction	0	L M M M L	M	L
hpxqcpthp	426	construction	We need a theory in which symmetry breaking begets further symmetry breaking in a progressive	construction	of diversifying structures and processes .	L L M U U	L	U
fkolixhal	412	construction	After any necessary revisions , a final	construction	permit is issued .	M L L L L	L	L
inkugoupx	401	construction	DNA- and peptide-based vaccines have become popular due to the comparative ease of	construction	, the circumvention of potential safety issues around attenuated organisms , and for their potentially enhanced immunogenicity compared to heat-killed and subunit vaccines [20 21 22] . DNA vaccines have the added attraction of efficiently priming both humoral and cytotoxic cell responses , a property largely lacking in subunit and attenuated organism vaccines .	M L L L L	L	L
qkodcuzzq	453	construction	A huge , fenced-off area is under	construction	as a zona franca - tax-free business development area - intended to attract foreign investment .	L L L M L	L	L
gzwmuqsgx	460	construction	Clinton will , says the paper , echoing reporting previously in the NYT , propose funding most of his social policy goals , from school	construction	to long-term care for the elderly , through targeted tax breaks rather than through new spending .	M L U L L	L	L
lpysalagc	495	damage	Microsoft is leveraging the power it has in the PC OS market in order to minimize the	damage	to its core business (the Windows software) .	U L L L M	L	U
kauufleya	474	damage	The toxicity produced by mutant SOD 1 that leads to mitochondrial	damage	remains to be identified .	M M L L M	M	M
aojyxzopl	489	damage	Fire did further	damage	in 1845 .	L M M M L	M	L
ibvyujkzb	481	damage	Because this was not in an urban area , missiles launched against it would have less risk of causing collateral	damage	0	M M M M M	M	M
mhpyxnwjm	400	classification	In a subject where there are no fixed criteria of definition , it is unfair to criticize an attempt at	classification	: after all , we have been putting up with ambiguous definitions for the eight parts of speech for as long as anyone can remember , yet we find their occasional application convenient .	L L L M L	L	L
fvednrtoz	193	instruction	For the Chief of Department 's	instructions	, see NYPD interview 8 , HQ (Feb .	M M M M L	M	M
vqyixvqxk	229	instruction	When the balloonist yells perfectly reasonable	instructions	to his would-be helpers in the field , they all ignore him because he appears incompetent .	M L M M L	M	M
kioeewphb	241	instruction	Having been only the second set of buildings erected in the country expressly for art	instruction	, Herron has had a long history of granting degrees through Indiana University for many of the state 's artists , art teachers , art scholars and visual designers .	M L L L M	L	L
pusqaflyw	237	acquisition	Ultimately , this should improve cost , schedule , and quality outcomes of DOD major weapon system	acquisitions	0	U M M L M	M	U

Table A.6: First twenty examples for ENG:PROCRES

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
jodkbnwa	1	glas	Tante Tyt fik fat i farmor Løjses	glas	og tømte det over i sit eget .	U L U	U	U
dkbuhvzl	2	dåse	Efter åbning skal indholdet hældes over i en anden beholder , hvis det skal gemmes , for at undgå , at der kommer bakterier i maden , og fordi smagen fra	dåsen	kan smitte af på indholdet , når der kommer ilt til .	L L U	L	U
dfvgjikf	3	dåse	Conns appetit lod til at være vendt tilbage til fuld styrke , men jeg følte mig så sammensnøret indvendig , at jeg ikke kunne få en bid ned , så jeg drak to	dåser	sodavand i stedet , mens jeg betragtede min bror spise kyllingeburger med pommes frites .	M M M	M	M
dpapszva	4	dåse	I maven på hajerne er fundet alt fra	dåser	til menneskelig og endda et krokodillehoved .	L M L	L	M
tuyfqaxw	5	glas	Der skulle åbenbart ikke så meget skum i , for han holdt	glasset	på skrå .	L M L	L	M
jeomutub	6	æske	En buket blomster , en	æske	fyldte chokolader eller en flaske vin går aldrig af mode , men med en anelse kreativitet kan den traditionelle erkendtlighed sagtens erstattes af en mere varig gave .	U U L	U	U
jhifnwbf	7	container	A.P. Møller overtager 70 containerskibe og 200.000	containere	med købet af Sea-Lands internationale containerdel , men analytikeren peger på , at en del af flåden formentlig bliver afhændet .	L L U	L	U
jzstowzl	8	flaske	Men det er en lang og træls vej , inden man kommer så vidt , at man kan flade ud i sofaen i sin nye stue , åbne en kølig	flaske	hvidvin og glemme alt om det forløbne halve års trængsler .	U L U	U	U
nmfrnrqv	9	fustage	Der skal også lige skiftes	fustager	på fadølsanlægget , og der er 117 andre ting at tage sig af .	L L U	L	U
abkvzkuv	10	gryde	Træk	gryden	fra varmen og lad risene stå 5 min .	L L U	L	U
jhsqeisz	11	kande	Fadet og	kanden	løb tilsammen op i 39,90 kr.	L L L	L	L
jdiszjep	12	kasse	For at understrege vigtigheden af et svar kan jeg oplyse , at der står et par	kasser	fluidum på spil .	U M M	M	M
jatnhzhv	13	krus	Men ordet dækker også over en bestemt type restaurant , som oprindeligt var et sted , hvor man ikke alene kunne drikker sig et	krus	øl fra fad , men også erhverve sig et enkelt og som oftest solidt måltid .	M M M	M	M
joberakl	14	pose	Det står altid på	posen	0	L L L	L	L
jzfcemij	15	skål	Bland alt i en	skål	og krydr forsigtigt med salt og rigeligt med peber .	L L L	L	L
jzptevwn	16	skrin	Dolly stod nøjagtig hvor han havde forladt hende , hun holdt	skrinet	i hænderne og stirrede på ham .	L L L	L	L
jjcsecmr	17	tønde	Polerede æbler Prinsen sloges den 17. februar 1946 med de andre kongelige børn fra Amalienborg og deres venner på Sorgenfri Slot om	tøndens	indhold af amerikansk slik , blankpolerede æbler og søde appelsiner .	L L L	L	L
jyhiuwfc	18	beholder	Affald indeholdende sådanne metalforbindelser skal naturligvis opsamles , hvilket sker i samme	beholder	0	L L L	L	L
jhuuyitc	19	container	Opel Millennium Express består af 42	containere	på 14 vogne og er dermed omkring 300 meter lang .	L U L	L	U
jarayilv	20	dåse	Fra hustage svirrede hvilende , stjålne Coca Cola	dåser	som små bomber ud over menneskemængden , hvor en del var ved at gå hjem , da Mandela begyndte sin tale .	L U L	L	U

Table A.7: First twenty examples for DA:CONTCONT

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
jcovqaos	1	Afghanistan	Det var først i den sidste time , at kaprerne udtrykte bekymring over tilstandene i	Afghanistan	.	M L L	L	M
jragnrqw	2	Afrika	"Rundrejsens vigtigste prestigeelement var dog USA's "" nye partnerskab "" med	Afrika	."	M M M	M	M
jcqtwujm	3	Amerika	"	Amerikas	kamp for at overvinde slaveriet , var et af de største slag i vores historie """" , fastslog Bill Clinton ."	M L M	M	M
jmpgicit	4	Århus	"Den første kontakt mellem Anders Nørgaard og politiet fandt sted i januar 1981 , hvor Nørgaard henvendte sig til Helsingør Politi med oplysninger om den vesttyske terrorgruppe Rote Arme Fraktion og en række "" politiske "" ildspåsættelser i	Århus	i 1980 ."	L L L	L	L
jdcwxgzj	5	Berlin	Tysk kraftcenter Talsmanden siger , at enhver diskussion om , at	Berlin	skulle kunne tage kræfter fra Frankfurt , er afsluttet .	M L M	M	M
jwnyaufi	6	Boston	Blot kan man se , at mængden af online-informationer er sendt af sted fra kraftfulde computere i New York , Chicago og	Boston	.	L L L	L	L
jhciamrt	7	Danmark	Denne måneds lønsedler skulle have set anderledes ud for størstedelen af	Danmarks	ca. 680.000 offentlige ansatte .	M L M	M	M
jdgshapu	8	England	End nu har den danske forsikringsbranche ikke besluttet sig for , om den vil indføre en generel undtagelsesklausul i alle forsikringspolicer , som det er tilfældet med kollegaerne i	England	.	M L L	L	M
jjtlbjtp	9	Europa	Ja-sigerne brugte argumenter om EUs udvidelse mod Østeuropa , Danmarks placering i	Europa	, konsekvenserne af et nej og fordelene ved Amsterdam-traktaten .	L L U	L	U
jeikkpbp	10	Frankrig	Markedsandelen er på 95 pct. , mens Skiold kun har en markedsandel på 1 pct. . Acemo er beliggende i Pontivy , Bretagne , som er det vigtigste område for svineproduktion i	Frankrig	.	L L L	L	L
jybqzjdm	11	Fyn	I Århus Amt tæller ventelisten mellem 160 og 170 børn , mens 110-120 børn konstant står på ventelisten på	Fyn	.	M L L	L	M
jtskanjn	12	Japan	Australien ,	Japan	og Canada har lagt op til at yde milliardbeløb i hjælp og kreditter til det nødstedte Indonesien for at afhjælpe følgerne af den seneste tørke og røgkatastrofe .	M M M	M	M
jpddluae	13	Jylland	Tyverier på Sjælland satte politiet igang Mens rambuktyvene før primært opererede i	Jylland	, er der i år begået flere af de voldsomme rambuktyverier på Sjælland .	L L L	L	L
jkmqpwjt	14	Kina	Dalai Lama inviterede i 1997 til forsoning med en erklæring om , at han er rede til at acceptere Tibet som en autonom del af	Kina	.	U M U	U	U
jpvottzb	15	København	14 personer kunne forlade festen på Charlottenborg i	København	med hæder .	L L L	L	L
jjkqedms	16	London	" Jeg har ikke hørt nogen sætte spørgsmålstegn ved Storbritanniens EU-medlemskab , fordi en bombe eksploderer i	Londons	undergrundsbane """" , siger en lettisk diplomat ."	U L L	L	U
jstycvcr	17	Paris	Som et skræmmende eksempel fremhæves det , hvordan	Paris	er parat til at bryde alle fælles regler , når det gælder sager af stor national interesse .	M M M	M	M
jydcanws	18	Tyskland	Møbeldetailhandelen i	Tyskland	lider voldsomt under lavkonjunkturen .	M L L	L	M
jkgbwnsr	19	Afghanistan		Afghanistan	og den borgerkrig , der brød ud i kølvandet på kommunisternes kup i april 1978 , var et tilbagevendende tema på politbureauets møder i 1979 .	M U U	U	U
jitmeykq	20	Afrika	Beslutningen kom som en total overraskelse for såvel sagsøgeren Paula Jones som præsidenten , der var ved at runde sin 12-dages succesfulde rundrejse i	Afrika	af med et besøg i Senegal .	L L L	L	L

Table A.8: First twenty examples for DA:LOCORG

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
psupazswk	1	barril	El precio del	barril	de brent del mar del Norte subió ayer hasta los 53,20 dólares , el máximo histórico , y cerró en 52,84 .	U M L U M M M M	M	M
bigawhlge	2	bol	Encontramos el	bol	boca abajo y encajado en un bloque de barro , explican en Nature los arqueólogos .	L L L L L L L L	L	L
prnntkwyb	3	botella	En julio de 1987 la policía de Daytona Beach detuvo a Moore y Susan Blahovec (Wuornos) para ser interrogadas bajo sospecha de haber golpeado a un hombre con una	botella	de cerveza .	L L L L L L L L	L	L
mezcoojfm	4	botellín	La ingeniería Compass adelgaza los	botellines	de Heineken	L L U L L L U L	L	L
veqhztakm	5	cajetilla	No nos gusta esta estrategia de bajos precios , agregan , porque irrita a las autoridades sanitarias y fiscales y reduce nuestros márgenes y los de la red minorista de distribución que trabaja con un porcentaje fijo sobre el precio de venta al público de las	cajetillas	.	U M U U U M M U	U	U
krbbicnvo	6	contenedor	Lo habitual es que estén llenos a rebosar , de manera que a los usuarios no les queda otro remedio que dejar sus desechos de reciclaje , ya sean plásticos , cartones o botellas , en bolsas de plástico y amontonados fuera de los	contenedores	.	L L L L L L L L	L	L
crdzyjflv	7	envase	En un intento de contrarrestar esta sensación , el 64 % lee detenidamente las etiquetas y nueve de cada diez vigilan el estado de los	envases	, se fija en la fecha de caducidad , sigue las instrucciones de conservación y mantiene la cadena del frío .	M U L U L L U L	L	U
lvbouwixc	8	lata	En los grandes supermercados se puede ver ya la mousse de avestruz en	latas	de 95 y 200 gramos , que se obtiene de triturar y sazonar el hígado del animal .	U L L L L M U L	L	L
dvagklnty	9	olla	Para ello el viejo Inca debió beber una	olla	llena de chicha hasta dejar la vacía .	M M M M M M M M	M	M
czoamblrc	10	taza	Si eres inglés , lo mismo que si eres chino , siempre necesitas una	taza	de té en esos momentos .	M M M U M M M M	M	M
dsyilklom	11	tazón	Divina Flor , su hija , que apenas empezaba a florecer , le sirvió a Santiago Nasar un	tazón	de café cerrero con un chorro de alcohol de caña , como todos los lunes , para ayudar lo a sobrellevar la carga de la noche anterior .	M M U U U M M M	M	M
pewuwgrxx	12	barril	Por eso , las últimas estimaciones de la OPEP para un precio del crudo más sostenido lo situaría entre los 22 y los 28 dólares por	barril	.	U M L U M M M U	M	M
rfkuxgzgj	13	bol	Los	boles	con sopa o arroz se sirven de forma individual .	M U M L U M L M	M	M
yqmkswdqb	14	botellín	La policía se limita a cortar la calle para que los bomberos , que vienen detrás , apaguen el fuego del contenedor y a recoger los restos de varios	botellines	de cerveza utilizados por los jóvenes para fabricar los cócteles incendiarios .	L L L L L U L U	L	L
jlzeorhxm	15	cesto	Un renglón importante de su economía es la artesanía la cual está representada en tejidos de palma de iraca ; los trabajos de este vegetal atraen a los turistas para comerciar sombreros , carteras , paneras ,	cestos	, muebles y adornos varios .	L L L L L L L L	L	L
qywrcenfn	16	envase	Los cigarrillos de este lanzamiento de la tabaquera estadounidense , el último en años , son más cortos y gruesos y contienen una nueva mezcla de tabaco , mientras que la cajetilla se presenta en un	envase	plateado con abertura lateral a modo de encendedor .	L L L L L L L L	L	L
bdxyvjkvc	17	frasco	Hoffman y Amanda colocan a Kerry en la Trampa del Angel y a su lado dejan colgado un	frasco	con ácido y una llave .	U L L L L L L L	L	L
tnqicunes	18	paquete	Los	paquetes	contienen 43 monedas de euro de diferente valor : cuatro de un céntimo , nueve de dos céntimos , seis de cinco céntimos , otras seis de 10 céntimos , siete de 20 céntimos , otras siete de 50 céntimos , dos piezas de euro y dos más de dos euros .	L L L L L L L L	L	L
jhimpafzl	19	taza	A partir de la	taza	ochocientos mil todo fue coser y cantar	U U M L L U M U	U	U
hrvwoyuvf	20	barrica	El aparato lo ha desarrollado un equipo de once investigadores de la Universidad de Valladolid y tiene un tamaño similar al de un microondas doméstico , en el que se introduce una muestra de un vino y el catador ofrece en un tiempo de unos quince minutos una completa información , que abarca desde el tipo de uva con el que está elaborado hasta el tiempo que ha estado en una	barrica	de roble .	L L L L L L L L	L	L

Table A.9: First twenty examples for SPA:CONTCONT

sentence	id	lemma	leftcontext	headword	rightcontext	annotations	vote	mace
wpqhrmami	1	Afganistán	Cuando crecí , empecé a ser maestra en otras escuelas clandestinas en	Afganistán	y en los campos de refugiados para alfabetizar mujeres , y después ingresé en el comité ejecutivo de la RAWA .	L L L L L L	L	L
oglxdoowl	2	África	Respecto a lo primero , baste recordar que el 12 % de la población del mundo , el que vive en Norteamérica y Europa occidental , es responsable del 60 % del consumo total , mientras que el 33 % de la humanidad , el que vive en el sur asiático y el	África	subsahariana , representa el 3,2 % del consumo global .	L L L U L L	L	L
cyanznjoc	3	Alemania	Lippi y Klinsmann dejan los banquillos de Italia y	Alemania	tras el Mundial	M M M M M M	M	M
ipfkdlgvl	4	Argentina	Entre las 23 películas aspirantes del Oso de Oro y los nueve filmes que se proyectarán fuera de concurso , figuran también películas de Hungría , Irán y	Argentina	.	L M U M L M	M	M
kbcmwnnqx	5	Cataluña	Pues bien , quien esto escribe fue diputado constituyente y ve más sintonía en el espíritu que animan las actuales reformas estatutarias , y no sólo la de	Cataluña	, con el clima y el ambiente que animó la elaboración de la Constitución del 78 , que lo que ahora dicen , con veinticinco años más , algunos de los protagonistas de entonces .	M M L M U U	M	M
iovyobmaq	6	China	Se distribuyen por	China	, Kazakhstan , Corea , Mongolia y Siberia .	L L L L L U	L	L
hjulmjhkm	7	España		España	, que trabaja intensamenteen el proceso de unión europea , entiende que ese proceso sólo debe avanzar manteniendo y mejorando la relación atlántica .	M M M M M M	M	M
dlxvaewlc	8	Europa	Habrá quienes cándidamente piensen que la antropología nació en los Estados Unidos o en algún país del norte de	Europa	.	L L L L L L	L	L
zxoaveeag	9	Francia	Las informaciones hacían referencia a que la NSA escuchó las conversaciones telefónicas de la princesa sin la aprobación de los servicios secretos británicos , la noche de su fallecimiento en París (	Francia	) el 31 de agosto de 1997 .	L L L L L L	L	L
hfpussrmd	10	Japón	Viaja frecuentemente para dictar clases magistrales de cocina peruana a diferentes países como	Japón	, China , Malasia , India , Holanda , España , Centroamérica , y Sudamérica , vocación y pasos que han empezado a seguir tres de sus cinco hijos .	L L L L L L	L	L
qjxycamec	11	Londres		Londres	es un hervidero de rumores respecto al porqué de esa permisividad , hasta el punto de que Whitehall ha tenido que desmentir que sea un agente doble (igual que el clérigo Abu Qatada , cuya extradición a EE .	M M U M U U	M	M
sozpcceuy	12	México	Para el ahora presidente de la Fundación para el Análisis y los Estudios Sociales (FAES) , que ha sido recibido en	México	por el presidente Felipe Calderón , durante el Gobierno del PP España estuvo junto a las democracias más importantes del mundo defendiendo una política que consistía , según él , en expandir la libertad y la democracia .	U L L U L U	L	U
hevnlhyjk	13	París	Alemania y Francia han incumplido durante varios años esa estipulación , pero , en contra de la voluntad de la Comisión Europea , el Consejo de Ministros de Finanzas (Ecofin) decidió en noviembre pasado suspender el procedimiento sancionador contra	París	y Berlín .	M M M M M M	M	M
riewujbel	14	Valencia	En los cuatro hospitales de	Valencia	que han recibido heridos quedaban anoche 10 personas .	L L L L L L	L	L
dwoptcomw	15	Afganistán	Pero también sabían que un final trágico del caso , en vísperas de un voto en el Parlamento sobre la renovación de la misión militar en	Afganistán	, iba a romper la coalición y lanzar el Gobierno en una crisis .	L L L U L L	L	L
wxafjtlhi	16	África	La gripe aviar irrumpe en	África	con la muerte de 40.000 aves en Nigeria	L L L L L L	L	L
mllfueavz	17	Alemania	Comparte su plataforma y mecánica con el Mercedes-Benz Clase E , y se fabrica en Sindelfingen ,	Alemania	.	L L L L L L	L	L
mpkwwsrki	18	América	Estaré en	América	de marzo a mayo , de junio a octubre en mi tierra (España) y de noviembre a diciembre me presentaré en Estados Unidos .	L L L U L L	L	L
qrjsvgkit	19	Argentina	En el 2007 inicia una gira por todo el Uruguay (350.000 personas en 32 conciertos) ,	Argentina	y por 7 ciudades de España ; la cual tiene su punto más alto en los conciertos de los días 24 , 25 , 26 , 27 , 28 y 29 de Julio en el teatro Teatro Solís de Montevideo a teatro lleno .	L L L L L L	L	L
pgkzfbczl	20	Cataluña	Regresó a	Cataluña	en 1970 .	L L L L L L	L	L

Table A.10: First twenty examples for SPA:LOCORG

Appendix B

Choosing dataset size

This appendix describes the preliminary study to determine the amount of data to annotate with turkers and volunters, which we set to 500 examples per dataset. The procedure consisted on evaluating the convergence of the accuracy of classifiers for three datasets annotated with expert annotation, using the features described in Martínez Alonso et al. (2012).

1. There are 1000 expert-annotated examples for the English datasets ENG:ARTINFO, ENG:CONTCONT and ENG:LOCORG.

2. We use the following features to characterize h for each dataset. These are the provisional features used in the first experiments (Martínez Alonso et al., 2011; Martínez Alonso et al., 2012) and have been replaced by the features in Chapter 6 for the experiments in this dissertation.

   (a) **NP-traits (6 features)**: these features describe the internal structure of the NP where t appears. The features indicate the presence of an adjective in the NP, of a common noun before or after t, of a genitive mark after t, of a coordinate "X and Y" and the presence of an article in the NP.

   (b) **Position of *t* (2 features)**: t being the first or last token of the sentence. This is a coarse approximation of the selected subject position for English (beginning of sentence) or for adjunts (end of sentence), as no parsing has ben used.

   (c) **Prepositions before *t* (57 features)**: each feature indicates whether the NP where t is included is introduced by a given preposition. The list of prepositions has been taken from the Preposition Project (Litkowski and Hargraves, 2005).

   (d) **Previous and next token after *t*'s NP (4 features)**: each feature describes whether the previous or next token is either a comma or a parenthesis.

   (e) **Verb after of before *t* (4 features)**: informs whether there is a verb immediately before t, or whether there is a modal or non-modal verb thereafter.

   (f) **Lexical space (3000 features)**: A bag of words with 3000 most frequent content words from the ANC.

3. Train and test the K-Nearest Neighbors (K=7) and Decision Tree classifiers—the highest performing for this data and annotation—and trace learning curves of accuracy for increments of 100 instances using 10x10 CV.

4. Find amount of training data where classifiers converge. We determine that classifier behaviors stabilize after 400 examples and we set the annotation task for 500 examples per dataset.

5. The blue line in the tables shows the evolution of accuracy as the number of training examples increases, and the green line shows the variation in standard deviation for the accuracy in each run.

6. Note that the additional expert annotations for not incorporated into the study in Chapter 4, we have only used them to calculate classifier convergence.

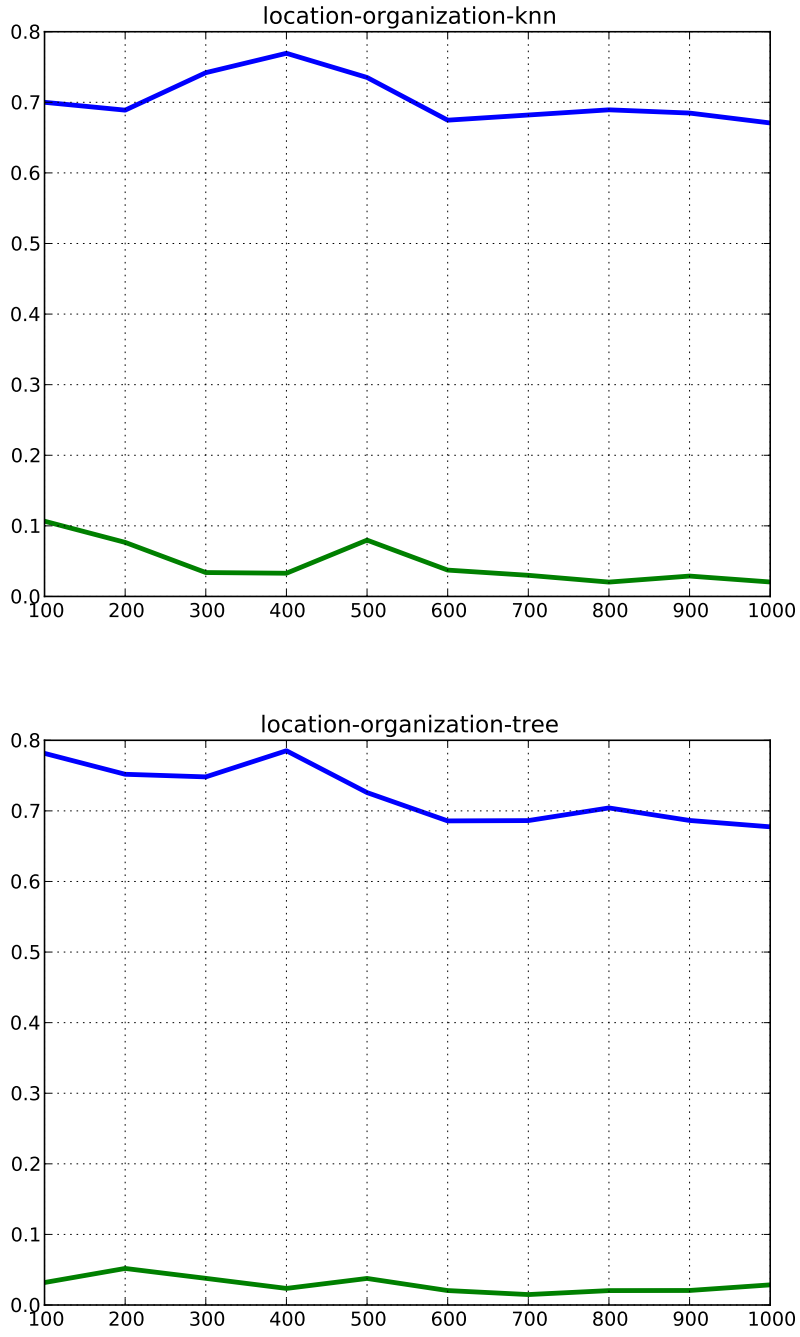

Figure B.1: Classifier convergence for ENG:LOCORG

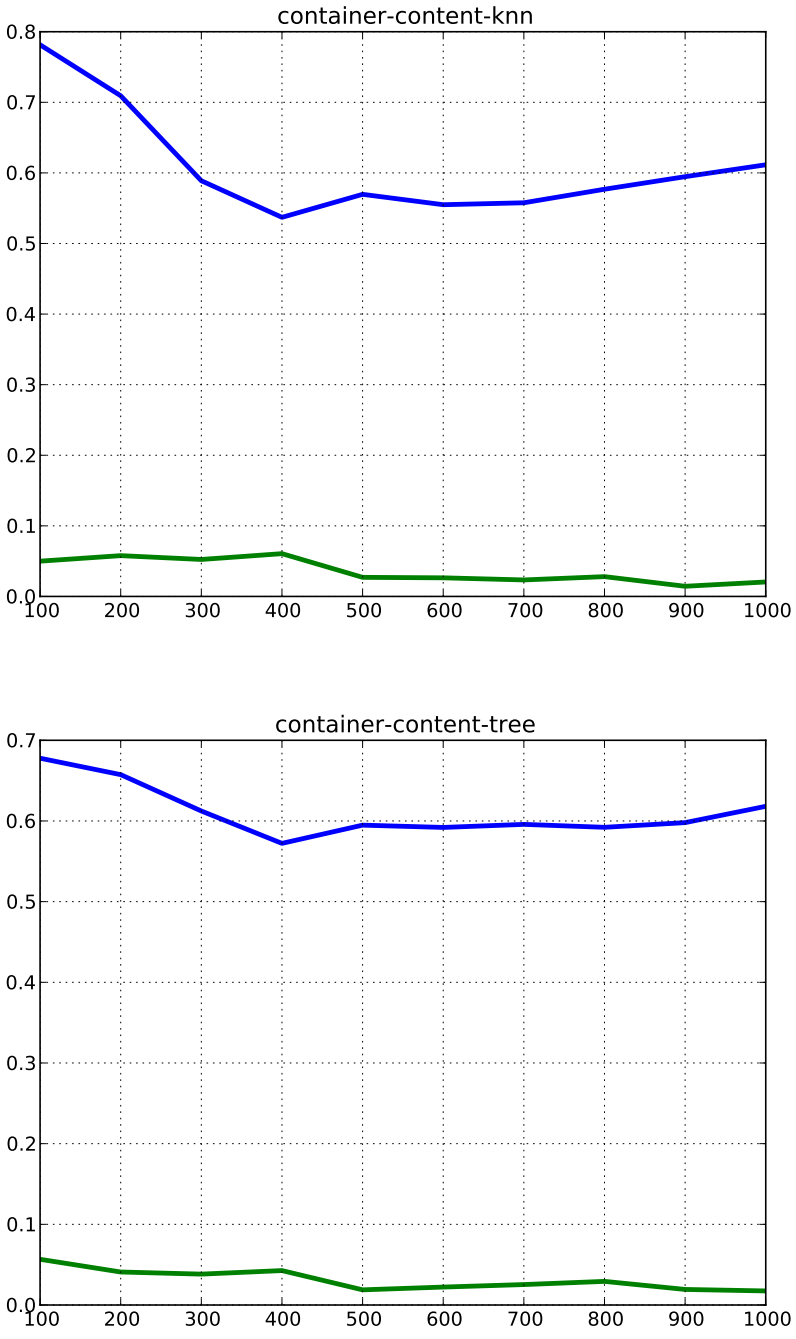

Figure B.2: Classifier convergence for ENG:CONTCONT

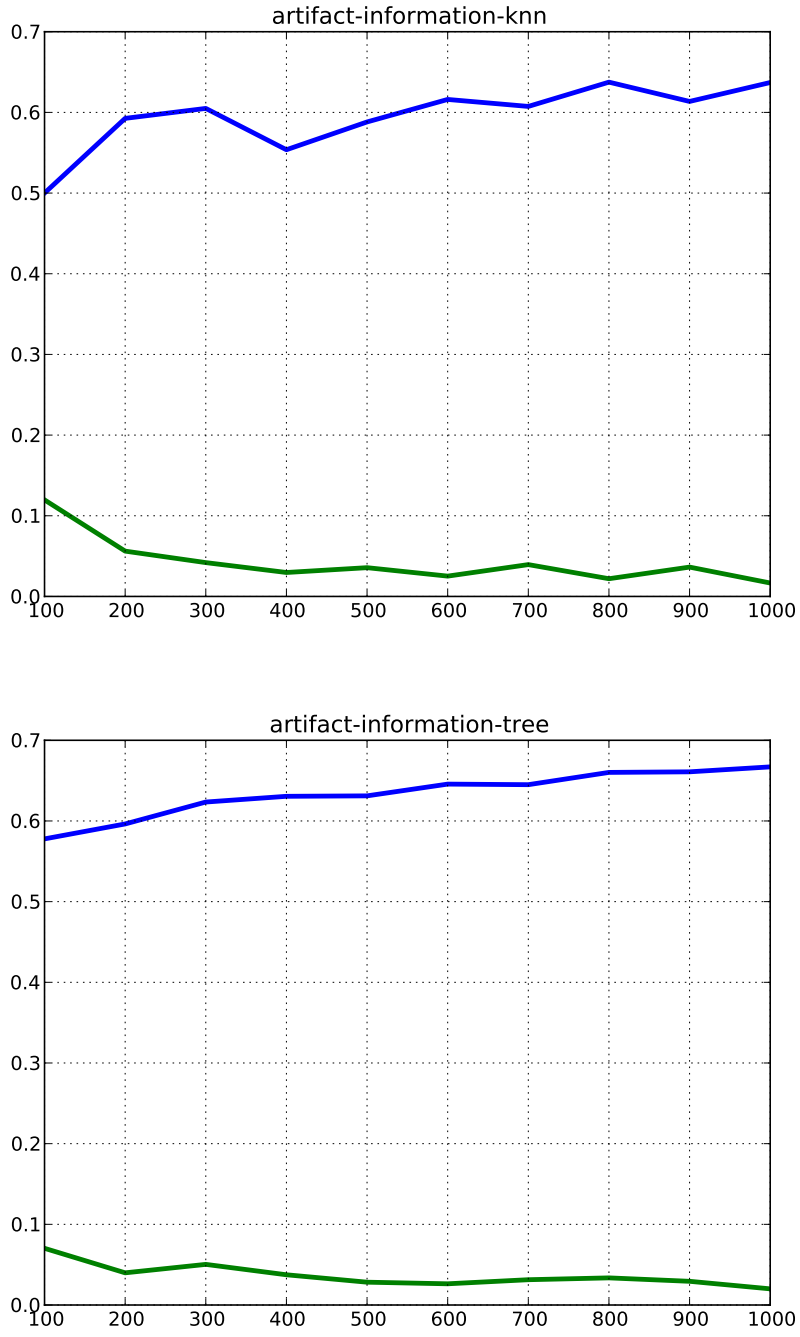

Figure B.3: Classifier convergence for ENG:ARTINFO

Appendix C

Full tables for WSI evaluation

C.1 WSI evaluation for mace

Table C.1 shows the evaluation for WSI on MACE for all datasets values of K.

dataset	k	description	hom	com	v-me
animeat	2	mace	0.0012480782	0.0013532291	0.0012985284
dk:contcont	2	mace	0.0109064933	0.0171545206	0.0133349183
artinfo	2	mace	0.005385482	0.0086797353	0.0066468306
dk:locorg	2	mace	0.0441326432	0.0669615362	0.0532015196
contcont	2	mace	0.0046377988	0.0058903612	0.0051895697
es:locorg	2	mace	0.0127478643	0.0165405237	0.0143986314
es:contcont	2	mace	0.0044200592	0.0064545692	0.0052469982
locorg	2	mace	0.0065615509	0.0076535258	0.0070655966
procres	2	mace	0.0025525213	0.004009777	0.0031193466
dk:contcont	3	mace	0.0046388278	0.0048708153	0.0047519919
procres	3	mace	0.0033344054	0.0034685489	0.0034001546
es:locorg	3	mace	0.0085109002	0.01004118	0.0092129271
artinfo	3	mace	0.0169040433	0.0183614978	0.0176026537
contcont	3	mace	0.0256370309	0.0198008568	0.0223441363
dk:locorg	3	mace	0.03486176	0.0508972385	0.0413803179
locorg	3	mace	0.1076258863	0.0781509358	0.090550195
animeat	3	mace	0.0098950368	0.0065085295	0.0078522119
es:contcont	3	mace	0.0045793679	0.0043766375	0.0044757081
es:locorg	6	mace	0.002087219	0.0013488931	0.0016387331
dk:contcont	6	mace	0.0101028047	0.0081090969	0.0089968224
procres	6	mace	0.0087756569	0.0053920672	0.0066798212
dk:locorg	6	mace	0.0166130643	0.0156449417	0.0161144754
artinfo	6	mace	0.0193493527	0.0130189756	0.0155651381
animeat	6	mace	0.0304043448	0.0123719509	0.0175873603
contcont	6	mace	0.0863700788	0.0433121679	0.0576929449
locorg	6	mace	0.1056682788	0.0513482123	0.0691121955
es:contcont	6	mace	0.0140136906	0.011832614	0.0128311257

Table C.1: MACE: Results of clustering solutions for each class in terms of homogeneity (HOM), completeness (COM) and V-measure (V-ME)

C.2 WSI evaluation for vote

Table C.2 shows the evaluation for WSI on MACE for all datasets values of K.

dataset	description	k	hom	com	v-me
eng:animeat	vote	2	0.0031256181	0.003063542	0.0030942687
dk:contcont	vote	2	0.0240587467	0.0311458476	0.0271473803
artinfo	vote	2	0.0094217442	0.0124454354	0.0107245388
dk:locorg	vote	2	0.073702729	0.0972331337	0.0838483767
contcont	vote	2	0.0068932903	0.0076767407	0.0072639519
es:locorg	vote	2	0.0142930784	0.0179458959	0.015912547
es:contcont	vote	2	0.0024624142	0.0034652362	0.002878998
locorg	vote	2	0.0014801147	0.0017504825	0.0016039851
procres	vote	2	0.0005058806	0.0006583682	0.0005721383
dk:contcont	vote	3	0.008016889	0.0069284007	0.0074330067
procres	vote	3	0.0049208316	0.0042407036	0.0045555222
es:locorg	vote	3	0.0104541333	0.0119350661	0.0111456215
artinfo	vote	3	0.0212949183	0.0189579061	0.0200585707
contcont	vote	3	0.0290482413	0.0196723565	0.0234581423
dk:locorg	vote	3	0.0385886461	0.0489857427	0.043170007
locorg	vote	3	0.1066620752	0.0785298566	0.0904592051
animeat	vote	3	0.0054957533	0.0032677671	0.0040985451
es:contcont	vote	3	0.0013307987	0.0012256881	0.0012760826
es:locorg	vote	6	0.0044963834	0.0028119033	0.0034600163
dk:contcont	vote	6	0.0227288786	0.0150155847	0.0180841041
procres	vote	6	0.0160742795	0.0081823335	0.010844475
dk:locorg	vote	6	0.0211670366	0.017332033	0.0190585269
artinfo	vote	6	0.0252433893	0.0139205027	0.0179451352
animeat	vote	6	0.0375477966	0.0138116448	0.0201948003
contcont	vote	6	0.1003335505	0.0441177214	0.0612869319
locorg	vote	6	0.1090561339	0.0537326335	0.0719935825
es:contcont	vote	6	0.0092577132	0.007532939	0.0083067397

Table C.2: VOTE: Results of clustering solutions for each class in terms of homogeneity (HOM), completeness (COM) and V-measure (V-ME)

C.3 WSI evaluation for expert

We have also used the expert annotations as test data to compare the learnability of the expert annotations with the two SAMs. We find that Expert fares either much better or much worse with regards to MACE and VOTE, the differences with the two SAMs are always significant. For instance, the expert annotations do worst for $k = 3$ in ENG:LOCORG, which is a highly stable dataset that barely changes sense assignments between MACE and VOTE (cf. Section 4.7.2). The four English datasets for VOTE that outperform MACE for $k = 3$ are ENG:ARTINFO, ENG:PROCRES, DA:CONTCONT, SPA:LOCORG. Three of them (excluding ES:LOCORG) are datasets where the difference between the output of both SAMs is greatest. This indicates that the more drastic updates taken by MACE are less reliable because the sense assignments correlate less with the distributional information. Drastic updates can be an artifact of applying EM to high-variance data (cf. Section 4.7.2).

dataset	k	description	hom	com	v-me
locorg	2	expert	0.0028909723	0.0042873501	0.0034533446
artinfo	2	expert	0.0032590015	0.0041925463	0.0036672957
contcont	2	expert	0.0034159931	0.0043836834	0.0038398086
procres	2	expert	0.0031177052	0.0048242296	0.0037876225
animeat	2	expert	0.0079832747	0.010608042	0.0091103728
locorg	3	expert	0.0038501875	0.0035545882	0.0036964877
contcont	3	expert	0.0037162248	0.0029000829	0.003257817
artinfo	3	expert	0.0050497124	0.0043781996	0.0046900414
animeat	3	expert	0.0107908665	0.008698541	0.0096323908
procres	3	expert	0.0117139422	0.0120025788	0.0118565041
contcont	6	expert	0.0121305841	0.0061463954	0.008158828
locorg	6	expert	0.0165642781	0.0102339474	0.012651431
animeat	6	expert	0.0182816204	0.0091167881	0.0121663752
artinfo	6	expert	0.019752388	0.0106081925	0.0138032363
procres	6	expert	0.0190064653	0.0115032111	0.0143321994

Table C.3: Expert annotations: Results of clustering solutions for each class in terms of homogeneity (HOM), completeness (COM) and V-measure (V-ME)

C.4 VOTE vs MACE WSI Evaluation

Table C.4 shows the division of the performance of WSI for MACE over WSI for VOTE. We determine the significance of the difference if it is higher than 1.05 or lower than .95.

dataset	description	k	hom	com	v-me
animeat	proportions	2	0.3993060534	0.4417204322	0.419655999
dk:contcont	proportions	2	0.4533275767	0.5507803417	0.491204609
artinfo	proportions	2	0.5716013807	0.6974231953	0.6197777508
dk:locorg	proportions	2	0.5987925245	0.6886699378	0.6344967156
contcont	proportions	2	0.6727989996	0.7672997384	0.7144278708
es:locorg	proportions	2	0.8918907381	0.9216883768	0.9048602554
es:contcont	proportions	2	1.7950104041	1.8626635363	1.8225084333
locorg	proportions	2	4.4331367147	4.3722379659	4.405026403
procres	proportions	2	5.0456995038	6.09047799	5.452084465
dk:contcont	proportions	3	0.5786319041	0.7030215991	0.6393095094
procres	proportions	3	0.6776101431	0.8179182409	0.7463808658
es:locorg	proportions	3	0.8141182076	0.8413175035	0.8265960802
artinfo	proportions	3	0.7938064424	0.9685403924	0.8775627145
contcont	proportions	3	0.8825674047	1.0065320239	0.952510902
dk:locorg	proportions	3	0.9034201379	1.0390214718	0.9585432298
locorg	proportions	3	1.009036118	0.9951748182	1.0010058661
animeat	proportions	3	1.8004878238	1.9917360455	1.9158534801
es:contcont	proportions	3	3.4410672355	3.5707595071	3.5073812447
es:locorg	proportions	6	0.4641995262	0.4797082081	0.4736200616
dk:contcont	proportions	6	0.4444919982	0.5400453607	0.4974989264
procres	proportions	6	0.5459440298	0.658988926	0.6159653839
dk:locorg	proportions	6	0.784855466	0.902660509	0.8455257615
artinfo	proportions	6	0.7665116778	0.9352374603	0.8673736885
animeat	proportions	6	0.8097504394	0.8957623132	0.8708855756
contcont	proportions	6	0.8608294871	0.9817408181	0.9413580209
locorg	proportions	6	0.9689347585	0.9556243379	0.9599771692
es:contcont	proportions	6	1.5137313422	1.5707831935	1.544664468

Table C.4: MACE over VOTE

Table C.5 shows the division of the performance of WSI for expert over WSI for VOTE. We determine the significance of the difference if it is higher than 1.05 or lower than .95.

dataset	description	k	hom	com	v-me
artinfo	expertvote	2	0.3459021384	0.3368742197	0.341953697
contcont	expertvote	2	0.4955533532	0.5710344475	0.5286115062
locorg	expertvote	2	1.9532082628	2.4492391234	2.1529779846
animeat	expertvote	2	2.5541427251	3.4626722879	2.9442733007
procres	expertvote	2	6.1629273366	7.3275555496	6.6201166141
locorg	expertvote	3	0.0360970616	0.0452641621	0.0408635876
contcont	expertvote	3	0.1279328671	0.1474191903	0.1388778777
artinfo	expertvote	3	0.2371322719	0.2309432067	0.233817328
animeat	expertvote	3	1.9634917963	2.6619219684	2.3501975859
procres	expertvote	3	2.3804801769	2.830327177	2.6026662958
contcont	expertvote	6	0.1209025699	0.1393180609	0.1331250852
locorg	expertvote	6	0.1518876339	0.1904605579	0.1757299828
animeat	expertvote	6	0.4868892984	0.6600798242	0.602450881
artinfo	expertvote	6	0.7824776509	0.7620552718	0.7691909894
procres	expertvote	6	1.1824147558	1.4058594776	1.3216130315

Table C.5: Expert over VOTE

Table C.6 shows the division of the performance of WSI for EXPERT over WSI for MACE. We determine the significance of the difference if it is higher than 1.05 or lower than .95.

dataset	description	hom	com	v-me
locorg	expertmace	0.4405928327	0.5601797392	0.4887548422
artinfo	expertmace	0.6051457364	0.4830269799	0.5517359998
contcont	expertmace	0.7365548307	0.7442130095	0.7399088527
procres	expertmace	1.2214217934	1.2031166621	1.2142358866
animeat	expertmace	6.3964538069	7.8390584535	7.0159209152
locorg	expertmace	0.0357738052	0.0454836289	0.0408225256
contcont	expertmace	0.1449553501	0.1464624938	0.1458018773
artinfo	expertmace	0.2987280768	0.2384445796	0.2664394512
animeat	expertmace	1.0905332268	1.3364833028	1.2267105028
procres	expertmace	3.5130527504	3.4604035409	3.4870485233
contcont	expertmace	0.1404489179	0.1419092069	0.1414181239
locorg	expertmace	0.1567573385	0.1993048422	0.1830564188
animeat	expertmace	0.6012831543	0.7368917116	0.6917681241
artinfo	expertmace	1.0208293932	0.8148254365	0.8868046144
procres	expertmace	2.1658168078	2.1333582736	2.145596272

Table C.6: Expert over MACE

Appendix D

Discussion on WSD baselines

D.1 Baselines

There are four baselines for this task. Three of them (MFS, SBJ,GRAM) are canonical baselines taken from the evaluation of the SemEval2007 metonymy resolution task in (Markert and Nissim, 2009). None of these three features uses a learning algorithm to make sense predictions. We include a fourth baseline (BOW) to determine the contribution of the feature sets to the learning algorithm. In this section we describe the four baselines and evaluate the three of them that do not require a learning algorithm.

D.1.1 MFS baseline

The most-frequent sense (MFS) baseline is a common baseline for WSD experiments. This baseline consist in assigning all the examples the most frequent sense, thus yielding perfect recall for the most frequent sense and zero to the other senses. Even though it is conceptually simple, MFS can be a hard baseline because sense distributions are often very skewed (McCarthy et al., 2004).

D.1.2 SBJ baseline

Like MFS, the subject baseline (SBJ) is another baseline that does not require a learning algorithm to assign senses. It does however require the examples to be annotated with the syntactic role of the headword. In SBJ, we assign the metonymic sense to any example where the headword is subject, and literal otherwise. This baseline has always zero recall for the underspecified sense.

This baseline was designed by Markert and Nissim (2009) for the SemEval metonymy resolution task. In that task, the metonymies could be often resolved by determining whether the headword is the subject of the sentence, because subjects are often agents, which organizations are more likely to be than locations.

D.1.3 GRAM baseline

The GRAM baseline is a supervised version of the SBJ baseline. Instead of assigning automatically the metonymic sense to the subject, it assigns the sense to each headword according to the most frequent sense of the syntactic role of the headword.

If subjects are indeed mostly metonymic in a dataset, headwords with the subject role will be tagged as metonymic. Nevertheless, this baseline is more free of assumptions in that it does not specify the relation between syntactic roles and senses beforehand. Also, this baseline does not automatically discard the underspecified sense.

However, this baseline is subject to the skewness of sense distributions, and the contribution of a more clearly defined role-sense relation (e.g. subject as often metonymic) can be outweighed by a more frequent, less informative role-sense association, like noun modifiers having the most frequent sense and taking up most of the probability mass for this feature.

D.1.4 BOW baseline

The BOW baseline is an additional fourth baseline we include to compare the contribution of the grammatical features and the more engineered semantic features with the contribution of a classifier trained only on a bag-of-words feature set, the `bow` feature set (cf. Table 6.8). Along with GRAM, this is another supervised baseline, but it is the only baseline that requires a classification algorithm.

D.1.5 Baseline comparison

In this section we provide an overview of the three baselines that do not require a classification algorithm, namely MFS, SBJ, GRAM. The BOW baseline is strictly speaking a feature set choice and we compare it with the classification system performance for other feature sets in Section 7.3.

Table D.1 shows the MFS, SBJ and GRAM baselines in terms of accuracy for the MACE and VOTE SAMs. Each dataset has two annotation variants with different sense distributions, and that changes the values of the baselines. The difference in sense distributions results in the lack of a unified baseline for both SAM.

DATASET	MACE			VOTE		
DATASET	MFS	SBJ	GRAM	MFS	SBJ	GRAM
ENG:ANIMEAT	.68	.60	.68	.72	.64	.72
ENG:ARTINFO	.34	.39	.37	.61	.40	.63
ENG:CONTCONT	.59	.54	.58	.71	.63	.71
ENG:LOCORG	.54	.59	.57	.54	.59	.57
ENG:PROCRES	.42	.41	.52	.69	.36	.59
DA:CONTCONT	.45	.42	.44	.66	.60	.65
DA:LOCORG	.50	.57	.60	.64	.68	.68
SPA:CONTCONT	.54	.54	.52	.58	.57	.58
SPA:LOCORG	.60	.68	.68	.63	.71	.71

Table D.1: Baselines for WSD

The MFS baseline is systematically lower for the MACE variant, with the exception of the ENG:LOCORG dataset. This difference in baseline is a consequence of the different way in which each SAM assigns senses to examples; MACE works globally and maximizes sense-likelihood given the input annotations, whereas VOTE operates locally and resolves ties with the backoff method described in Section 4.7.1. As an effect of the EM implementation of MACE, which has a bias for evenly-sized clusters, this SAM has a tendency to smooth the distribution of senses, while the resulting distributions of VOTE are more skewed.

The difference between the value of MFS is proportional to the difference coefficient D we use to measure the difference between annotation variants for each dataset, which we list in Table 4.12. Figure D.1 illustrates the difference in MFS for MACE and VOTE.

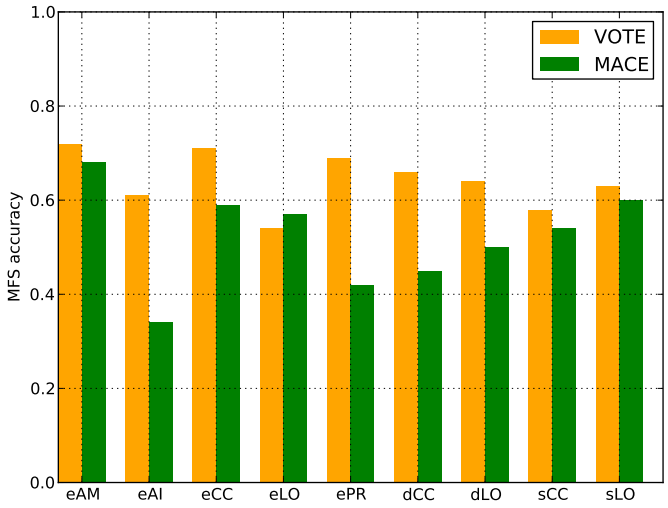

Figure D.1: Comparison of MFS for MACE and VOTE

With regards to the SBJ baseline, it is only systematically harder for all languages and annotation variants of the LOCATION•ORGANIZATION dot types. It does not provide a harder baseline for the rest of the datasets except ENG:ANIMEAT for MACE. This baseline was tailored by Markert and Nissim (2009) to assess the performance of metonymy resolution systems over the LOCATION•ORGANIZATION, but it falls short of a better, feature-based baseline for the rest of the datasets. We discard using SBJ as a reference baseline for the system.

The GRAM baseline offers a more varied behavior with regards to MFS. It only improves significantly (corrected paired t-test with $p < 0.05$) over four datasets for VOTE, while it fares worse in three and same in the other two. Out of the five datasets for MACE for which GRAM is a harder baseline than MFS, the GRAM value is lower than the baseline set by SBJ.

Even though the GRAM baseline aims at improving over SBJ and providing a more informed baseline that takes all senses and syntactic roles in consideration, it is also more affected by parser performance (subject is amongst the

easiest syntactic roles to identify, while other roles for nouns are less accurate), and by the skewness of sense distributions. Due to the unsystematic relation with MFS, we also discard GRAM as a comparison baseline for our system.

After comparing the behavior of the MFS, SBJ and GRAM baselines, we determine we will strictly use the MFS baseline to compare our classification experiments against. The SBJ baseline was specifically developed to evaluate WSD for the LOCATION•ORGANIZATION dot type, and is not a harder baseline for the other dot types. The GRAM baseline does not systematically improve over MFS. For these reasons we discard SBJ and GRAM from our analysis. In Section 7.3 we evaluate the WSD system against MFS and BOW, which a particular instance of a classification algorithm trained on the BOW feature set.

Appendix E

Tables for WSD

E.1 Evaluation tables for all datasets and classifiers for VOTE

The following nine tables provide the WSD evaluation for all datasets. Each table provides the overall acurracy for the dataset trained and tested on each combination for feature set and learning algorithm. Cf. Chapter 6.

The names of the columns stand for:

1. MFS: Most Frequent Sense baseline
2. NB: Naive Bayes classifier
3. LR: Logistic Regression classifier. The final system uses this classifier.
4. DT: Decision Tree classifier.

The bullets (•) represent significant improvement over the MFS baseline. The daggers (†) represent significant degradation over MFS. We calculate significance using a corrected paired t-test with $p < 0.05$.

Table E.1: WSD evaluation for ENG:ANIMEAT

Feature group	MFS	NB	LR	DT
ALB	0.72	0.72	0.84 •	0.77 •
SEM	0.72	0.79 •	0.80 •	0.70 †
PPW	0.72	0.76 •	0.78 •	0.71
PLE	0.72	0.72	0.73 •	0.70 †
PLB	0.72	0.76 •	0.81 •	0.78 •
PBR	0.72	0.72 •	0.75 •	0.73
ALL	0.72	0.72 •	0.82 •	0.72
TOP	0.72	0.72	0.72	0.72
BOW	0.72	0.77 •	0.81 •	0.78 •
PBW	0.72	0.72 •	0.80 •	0.74 •
BRW	0.72	0.78 •	0.77 •	0.68 †
WNT	0.72	0.76 •	0.77 •	0.71

Table E.2: WSD evaluation for ENG:ARTINFO

Feature group	MFS	NB	LR	DT
ALB	0.61	0.61	0.66 •	0.56 †
ALL	0.61	0.62 •	0.66 •	0.55 †
BOW	0.61	0.62 •	0.61	0.55 †
BRW	0.61	0.61	0.62	0.49 †
PBW	0.61	0.62 •	0.65 •	0.55 †
PBR	0.61	0.65 •	0.66 •	0.58 †
PLB	0.61	0.61	0.67 •	0.54 †
PLE	0.61	0.61	0.64 •	0.58 †
PPW	0.61	0.63 •	0.63 •	0.52 †
SEM	0.61	0.61	0.59 †	0.49 †
TOP	0.61	0.61	0.61	0.61
WNT	0.61	0.59 †	0.58 †	0.46 †

Table E.3: WSD evaluation for ENG:CONTCONT

Feature group	MFS	NB	LR	DT
ALB	0.71	0.71	0.83 •	0.80 •
ALL	0.71	0.73 •	0.82 •	0.74 •
BOW	0.71	0.72 •	0.74 •	0.67 †
BRW	0.71	0.69 †	0.73 •	0.60 †
PBW	0.71	0.75 •	0.84 •	0.76 •
PBR	0.71	0.78 •	0.84 •	0.78 •
PLB	0.71	0.71	0.74 •	0.66 †
PLE	0.71	0.71	0.70 †	0.65 †
PPW	0.71	0.72 •	0.71	0.62 †
SEM	0.71	0.72	0.74 •	0.66 †
TOP	0.71	0.71	0.71	0.71
WNT	0.71	0.73 •	0.72	0.63 †

Table E.4: WSD evaluation for ENG:LOCORG

Feature group	MFS	NB	LR	DT
ALB	0.61	0.61	0.73 •	0.68 •
ALL	0.61	0.64 •	0.72 •	0.67 •
BOW	0.61	0.64 •	0.63 •	0.58 †
BRW	0.61	0.60 †	0.62	0.53 †
PBW	0.61	0.67 •	0.71 •	0.63 •
PBR	0.61	0.70 •	0.73 •	0.74 •
PLB	0.61	0.67 •	0.73 •	0.70 •
PLE	0.61	0.67 •	0.72 •	0.72 •
PPW	0.61	0.70 •	0.70 •	0.63 •
SEM	0.61	0.62	0.61	0.54 †
TOP	0.61	0.61	0.61	0.61
WNT	0.61	0.60 †	0.59 †	0.49 †

Table E.5: WSD evaluation for ENG:PROCRES

Feature group	MFS	NB	LR	DT
ALB	0.60	0.60	0.60	0.55 †
ALL	0.60	0.61 •	0.60	0.52 †
BOW	0.60	0.60	0.58 †	0.51 †
BRW	0.60	0.56 †	0.56 †	0.47 †
PBW	0.60	0.61 •	0.62 •	0.57 †
PBR	0.60	0.62 •	0.65 •	0.61
PLB	0.60	0.60 •	0.60	0.55 †
PLE	0.60	0.61 •	0.60	0.55 †
PPW	0.60	0.59	0.57 †	0.51 †
SEM	0.60	0.58 †	0.55 †	0.49 †
TOP	0.60	0.60	0.60	0.60
WNT	0.60	0.57 †	0.55 †	0.44 †

Table E.6: WSD evaluation for DA:CONTCONT

Feature group	MFS	NB	LR	DT
ALB	0.66	0.66	0.58 †	0.56 †
ALL	0.66	0.65 †	0.57 †	0.54 †
BOW	0.66	0.65 †	0.61 †	0.55 †
BRW	0.66	0.59 †	0.57 †	0.52 †
PBW	0.66	0.65 †	0.61 †	0.54 †
PBR	0.66	0.65 †	0.64 †	0.56 †
PLB	0.66	0.66	0.60 †	0.55 †
PLE	0.66	0.66	0.64 †	0.54 †
PPW	0.66	0.61 †	0.60 †	0.50 †
SEM	0.66	0.62 †	0.55 †	0.51 †
TOP	0.66	0.66	0.66	0.66
WNT	0.66	0.64 †	0.64 †	0.48 †

Table E.7: WSD evaluation for DA:LOCORG

Feature group	MFS	NB	LR	DT
ALB	0.64	0.64	0.68 •	0.62 †
ALL	0.64	0.67 •	0.68 •	0.60 †
BOW	0.64	0.65	0.64	0.57 †
BRW	0.64	0.63 †	0.61 †	0.52 †
PBW	0.64	0.68 •	0.68 •	0.59 †
PBR	0.64	0.69 •	0.68 •	0.67 •
PLB	0.64	0.67 •	0.69 •	0.58 †
PLE	0.64	0.70 •	0.68 •	0.66 •
PPW	0.64	0.67 •	0.67 •	0.57 †
SEM	0.64	0.63 †	0.62 †	0.53 †
TOP	0.64	0.64	0.64	0.64
WNT	0.64	0.62 †	0.62 †	0.50 †

Table E.8: WSD evaluation for SPA:CONTCONT

Feature group	MFS	NB	LR	DT
ALB	0.58	0.60 •	0.73 •	0.64 •
ALL	0.58	0.69 •	0.70 •	0.56 †
BOW	0.58	0.69 •	0.70 •	0.62 •
BRW	0.58	0.67 •	0.66 •	0.53 †
PBW	0.58	0.64 •	0.65 •	0.54 †
PBR	0.58	0.62 •	0.64 •	0.56 †
PLB	0.58	0.67 •	0.71 •	0.64 •
PLE	0.58	0.59 •	0.60 •	0.52 †
PPW	0.58	0.60 •	0.63 •	0.52 †
SEM	0.58	0.66 •	0.67 •	0.55 †
TOP	0.58	0.58	0.58	0.58
WNT	0.58	0.59	0.59	0.48 †

Table E.9: WSD evaluation for SPA:LOCORG

Feature group	MFS	NB	LR	DT
ALB	0.63	0.64 •	0.75 •	0.67 •
ALL	0.63	0.73 •	0.76 •	0.67 •
BOW	0.63	0.63	0.64	0.56 †
BRW	0.63	0.65 •	0.65 •	0.58 †
PBW	0.63	0.71 •	0.73 •	0.61
PBR	0.63	0.70 •	0.71 •	0.70 •
PLB	0.63	0.71 •	0.72 •	0.64
PLE	0.63	0.70 •	0.71 •	0.69 •
PPW	0.63	0.72 •	0.72 •	0.60 †
SEM	0.63	0.65 •	0.65 •	0.56 †
TOP	0.63	0.63	0.63	0.63
WNT	0.63	0.63	0.63	0.50 †

E.2 Tables for the WSD system evaluated on MACE

From these differences we can see that each SAM yields a sense distribution that is best captured by particular feature sets. However, the feature-set ranking consistency across SAMs is not correlated with the difference in sense distributions. For instance, ENG:ARTINFO:MACE and ENG:ARTINFO:VOTE differ in about a third of their assigned senses (cf. Table 4.12), yet their rankings are very similar. The same applies for the heavily-updated dataset ENG:PROCRES, where the difference coefficient is 0.27 and the first three dataset for each SAM differ only in one, i.e. the sense tags in the ENG:PROCRES:MACE dataset are 27% different from ENG:PROCRES:VOTE.

The overall accuracy is always higher for VOTE, even though the MFS is also always highest for this SAM. All the accuracy scores for both SAMs are

Dataset	MACE	VOTE
ENG:ANIMEAT	alb, **all**, plb	alb, **all**, sem
ENG:ARTINFO	plb, alb, **all**	plb, **all**, alb
ENG:CONTCONT	pb, pbw,alb	pbr, pbw, alb
ENG:LOCORG	pbr, **all**, alb	plb, alb, ple
ENG:PROCRES	pbr, pbw, ple	pbr, pbw, alb
DA:CONTCONT	**all**, pbw, alb	pbr, pbw, wnt
DA:LOCORG	**all**, alb, plb	**all**, alb, plb
SPA:CONTCONT	**all**, alb, bow	alb, **all**, plb
SPA:LOCORG	**all**, alb, pbw	**all**, alb, pbw

Table E.10: Feature set performance ranking

	Acc-ALL	Acc-MFS	ER-MFS	Acc-BOW	ER-BOW
ENG:ANIMEAT	0.81	0.68	40.62%	0.79	9.52%
ENG:ARTINFO	0.47	0.34	19.70%	0.43	7.02%
ENG:CONTCONT	0.71	0.59	29.27%	0.63	21.62%
ENG:LOCORG	0.72	0.57	34.88%	0.63	24.32%
ENG:PROCRES	0.45	0.42	5.17%	0.37†	12.70%
DA:CONTCONT	0.4†	0.45	-9.1%	0.37†	4.76%
DA:LOCORG	0.6	0.5	20%	0.59	2.44%
SPA:CONTCONT	0.67	0.54	28.26%	0.67	0%
SPA:LOCORG	0.74	0.6	35%	0.63	29.73%

Table E.11: Accuracies and error reduction over MFS and BOW for MACE

statistically significant over their respective MFS except those marked with a dagger. However, overall accuracy blurs out the details of the performance over each sense.

	MACE			VOTE		
Dataset	L	M	U	L	M	U
ENG:ANIMEAT	.86	.68	.0	.88	. 68	.0
ENG:ARTINFO	.52	.48	.37	.54	.77	.02
ENG:CONTCONT	.77	.61	.02	.89	.69	.0
ENG:LOCORG	.79	.61	.0	.79	.61	.0
ENG:PROCRES	.41	.51	.25	.45	.71	.01
DA:CONTCONT	.50	.31	.35	.73	.14	.09
DA:LOCORG	.73	.55	.17	.82	.49	.14
SPA:CONTCONT	.78	.65	.22	.81	.64	.17
SPA:LOCORG	.85	.69	.06	.85	.68	.02

Table E.12: Sense-wise performance in terms of F1

	MACE		VOTE	
Dataset	Acc-L	Acc-M	Acc-L	Acc-M
ENG:ANIMEAT	0.81	0.84	0.83	0.84
ENG:ARTINFO	0.69	0.65	0.76	0.68
ENG:CONTCONT	0.69	0.75	0.83	0.86
ENG:LOCORG	0.75	0.73	0.75	0.74
ENG:PROCRES	0.64	0.58	0.7	0.61
DA:CONTCONT	0.53	0.69	0.58	0.8
DA:LOCORG	0.7	0.78	0.75	0.84
SPA:CONTCONT	0.73	0.79	0.74	0.83
SPA:LOCORG	0.8	0.82	0.8	0.83

Table E.13: Individual accuracies for classifiers

	MACE			VOTE		
Dataset	L	M	U	L	M	U
ENG:ANIMEAT	0.81	0.84	0.03	0.88	0.64	0.01
ENG:ARTINFO	0.69	0.65	0.43	0.47	0.75	0.15
ENG:CONTCONT	0.69	0.75	0.03	0.88	0.66	0.04
ENG:LOCORG	0.75	0.73	0.02	0.79	0.58	0.07
ENG:PROCRES	0.64	0.58	0.31	0.41	0.67	0.12
DA:CONTCONT	0.53	0.69	0.38	0.69	0.11	0.2
DA:LOCORG	0.7	0.78	0.32	0.81	0.45	0.25
SPA:CONTCONT	0.73	0.79	0.28	0.78	0.62	0.25
SPA:LOCORG	0.8	0.82	0.18	0.85	0.64	0.17

Table E.14: Sense-wise F1 scores for the ensemble system

The ensemble performs unevenly for each SAM. In the case of MACE, it improves the performance of the metonymic sense but alters the reliability of the literal sense. In VOTE, the ensemble performs in a similar manner to the single classifier, but the F1 is consistently lower for the alternating senses.

	MACE			VOTE		
Dataset	gold	averagedot/round	Excl.	gold	averagedot/round	averageexclusion
ENG:ANIMEAT	1.4	3.11	2.39	0.7	1.85	1.33
ENG:ARTINFO	15.0	22.53	21.71	5.4	9.67	8.79
ENG:CONTCONT	2.8	6.06	4.73	2.5	3.54	3.02
ENG:LOCORG	1.6	4.99	3.87	2.2	5.07	3.96
ENG:PROCRES	13.4	20.02	19.63	4.8	9.42	8.11
DA:CONTCONT	14.3	24.25	23.38	9.1	11.6	10.9
DA:LOCORG	10.5	14.39	13.81	8.3	10.62	10.11
SPA:CONTCONT	7.5	12.25	11.25	6.9	11.51	10.54
SPA:LOCORG	5.2	7.38	6.77	4.7	6.63	5.83

Table E.15: Run-wise amount of underspecified senses and amount of underspecified assigned by exclusion

	MACE			VOTE		
Dataset	p	r	F1	p	r	F1
ENG:ANIMEAT	.0	.0	.0	.0	.0	.0
ENG:ARTINFO	.41	.36	.37	.04	.01	.02
ENG:CONTCONT	.04	.01	.02	0	.0	.0
ENG:LOCORG	.0	.0	.0	.0	.0	.0
ENG:PROCRES	.30	.23	.25	.03	.0	.01
DA:CONTCONT	.37	.36	.35	.14	.08	.09
DA:LOCORG	.24	.15	.17	.20	.12	.14
SPA:CONTCONT	.30	.18	.22	.27	.15	.17
SPA:LOCORG	.12	.05	.06	.05	.01	.02

Table E.16: Performance for underspecified

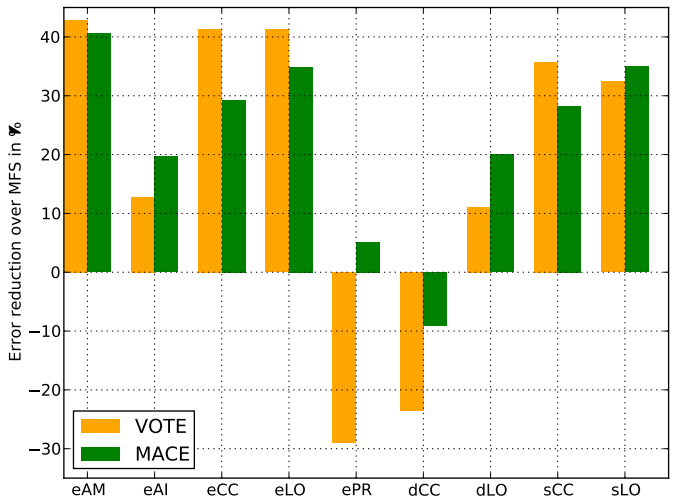

Figure E.1: Error reduction over MFS for MACE and VOTE

Appendix F

Literality prediction

F.1 Scatter plots for all datasets

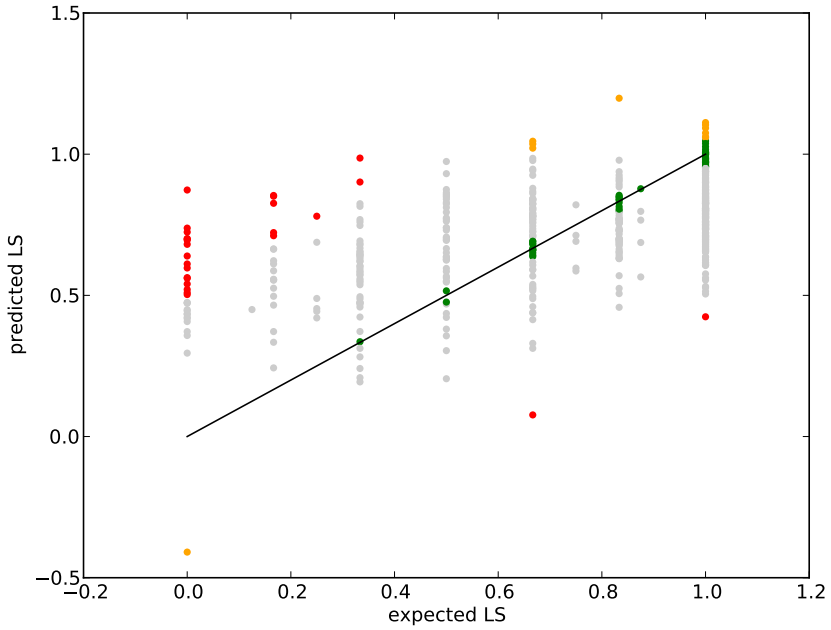

Figure F.1: Literality prediction scatter plot for DK:CONTCONT

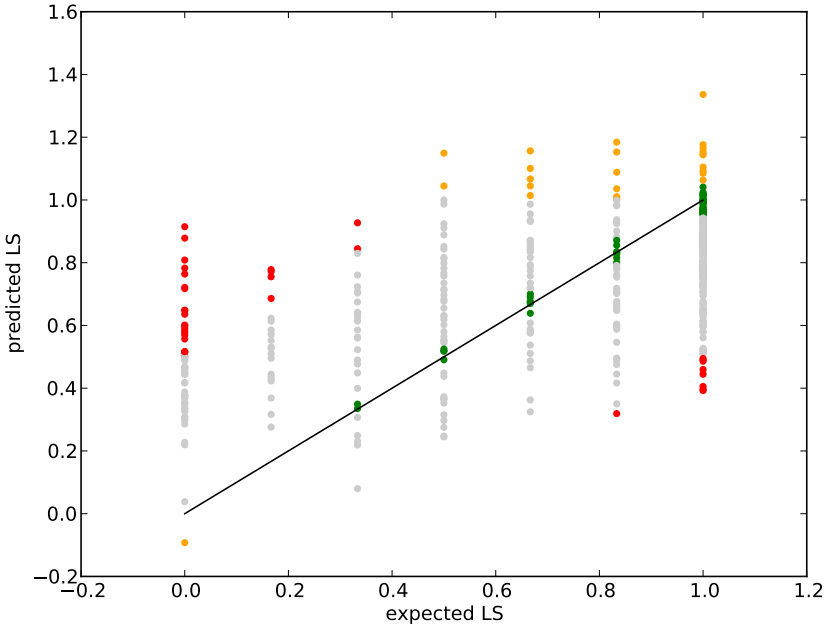

Figure F.2: Literality prediction scatter plot for DK:LOCORG

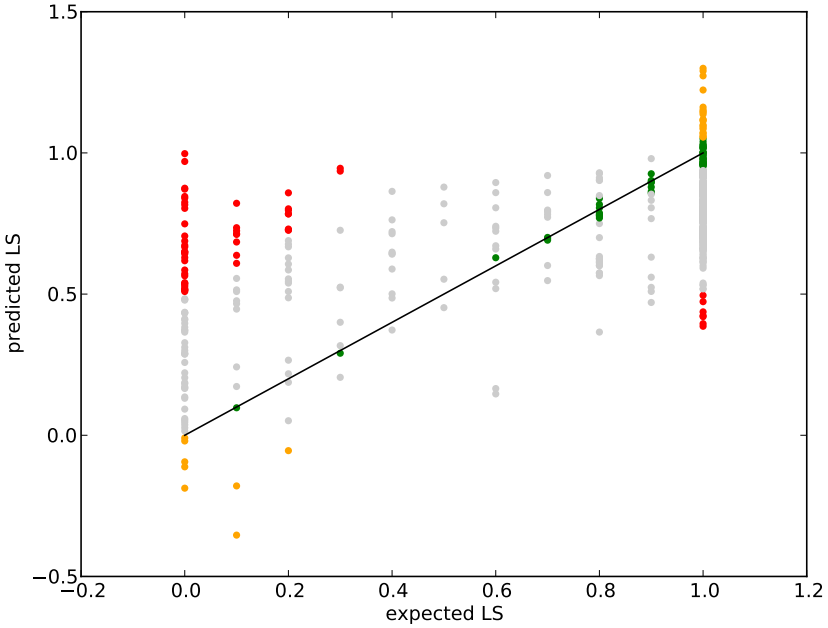

Figure F.3: Literality prediction scatter plot for ENG:ANIMEAT

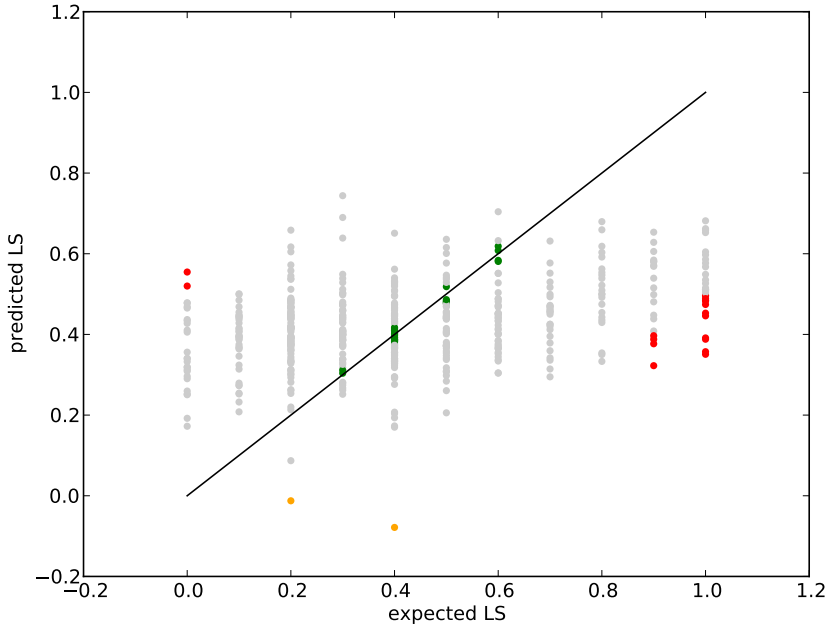

Figure F.4: Literality prediction scatter plot for ENG:ARTINFO

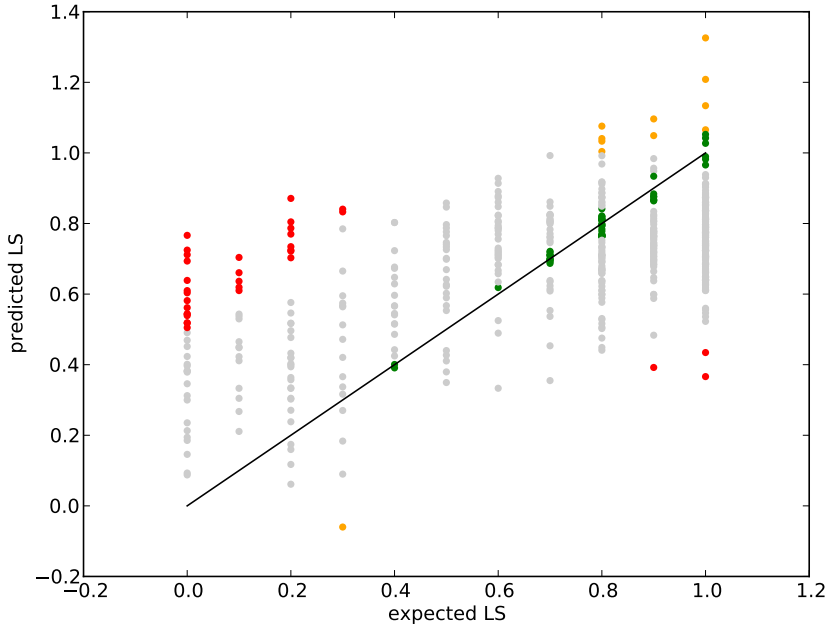

Figure F.5: Literality prediction scatter plot for ENG:CONTCONT

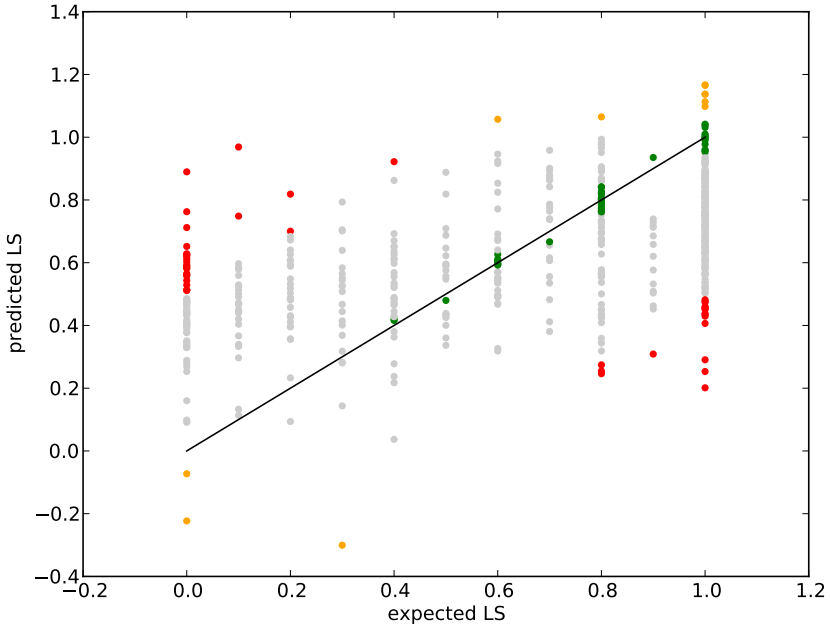

Figure F.6: Literality prediction scatter plot for ENG:LOCORG

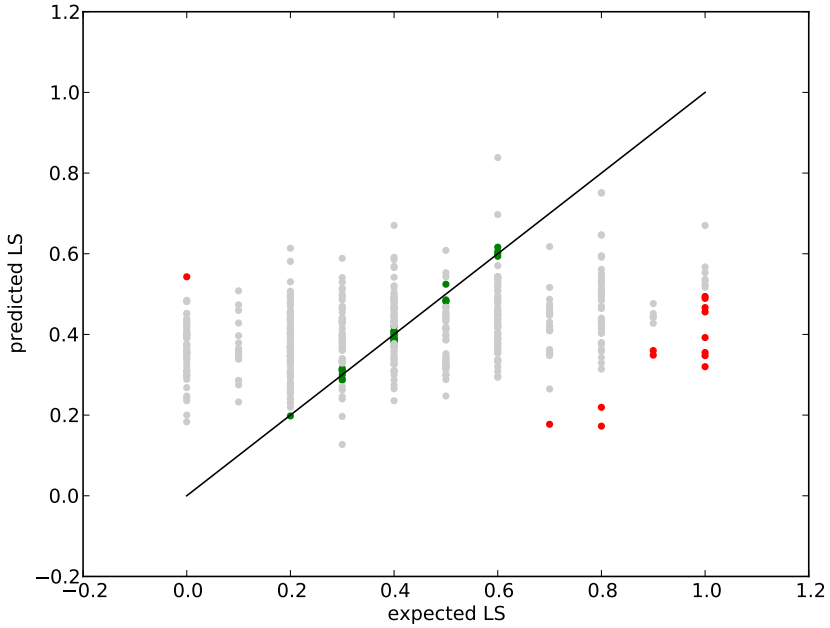

Figure F.7: Literality prediction scatter plot for ENG:PROCRES

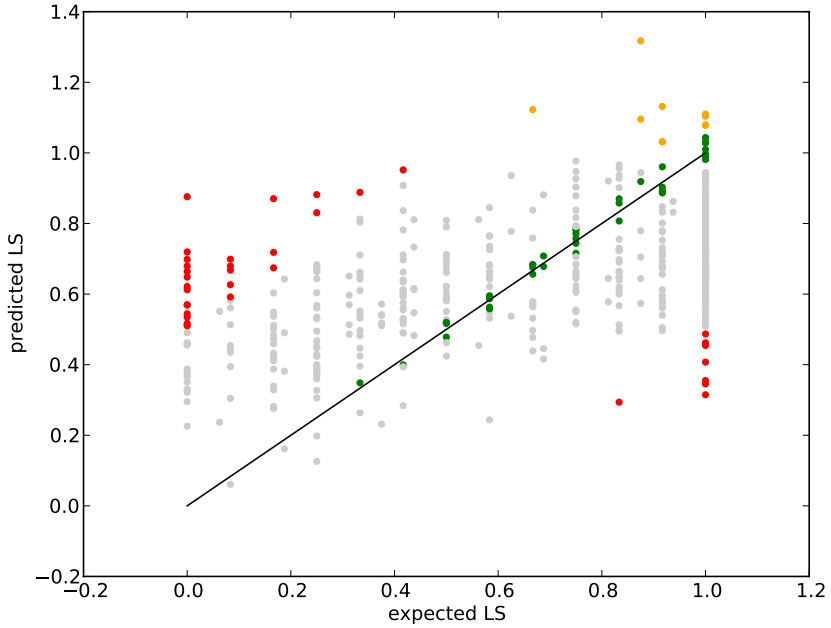

Figure F.8: Literality prediction scatter plot for SPA:CONTCONT

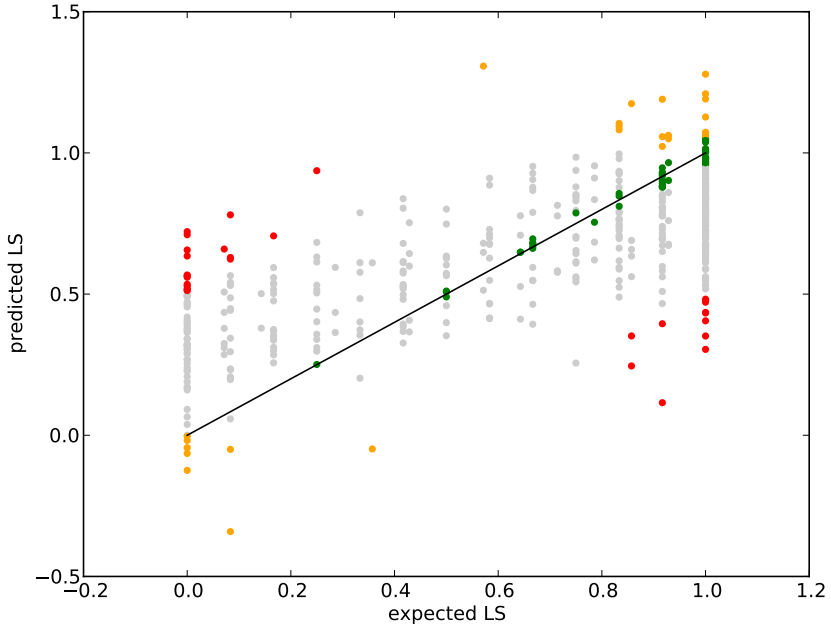

Figure F.9: Literality prediction scatter plot for SPA:LOCORG

F.2 High-coefficient features for literality

High positive coefficient	High negative coefficient
ENG:ANIMEAT	
bh:1100010, phw:or, w:n:feeling, bb:100001111111, bb:0000101001, bb:1111110101, bb:0011000100, bh:0001110000, phw:eat, bb:000101100, bb:00001011110, w:n:animal, bb:0000100110, bc:1001010111, bb:100111010, bb:00011101001, bb:111111111111110, w:v:consumption, bb:1111110100, w:n:location	m:plural, bb:00011100100, bb:011001110, w:v:contact, w:n:phenomenon, bb:1111111110, pl:sbj, bb:0010111100111, bb:10011100, w:n:motive, bb:0101010110, bb:11110000000, bb:0001111001, bb:01101010, bb:1111001111000, bb:111111111110, w:v:motion, w:v:social, bb:110110010, w:n:possession
ENG:ARTINFO	
bb:1111101011000, w:n:event, w:v:cognition, bb:1101101110, w:n:group, bb:10001100, bb:1111101101100, bb:101111010, bb:00001011100, bb:011011100, bb:0000101001, bb:011011001, bc:11010, bb:0000101000, bb:1111101100, bb:0111111, plc:nmod, bb:00001011101, bb:10011111001, pl:sbj	m:plural, bb:110110110, bb:00001011111, pl:coord, w:n:body, phw:and, bh:110000, bb:1111110101, plc:adv, bb:1111000110100, pl:obj, bb:0010010110, w:v:possession, bb:001010010110, bb:111111101100, bb:000010110110, bb:111110111111111, plh:nmod, bb:1111111011110110, bb:0001111011
ENG:CONTCONT	
bb:1111101010010, bb:11111110111001, t:38, bb:111101110, bb:001001001, plh:obj, bb:1111000110100, bb:1001010111, bb:0001110001, bb:1010111000, bb:11010, bb:1111110100, bb:11110011011011, bh:11000110, phw:per, bb:00011100110, bb:0101010110, bb:11101101011, w:n:location, bc:11010	bc:10001110, bb:00100110010, bc:00011101111, bb:1111111011111110, m:plural, bb:11111010101110, w:v:change, w:v:creation, bb:1101101111100, bb:0010111110000, bb:1111111000001, bc:0001110000, bb:0010111100100, bb:00110110100, bb:00011100101, plh:pmod, bb:000001100, bb:0000111110, bb:1111101111010, bb:111110111111101
ENG:LOCORG	
bb:00001011111, bb:0001101101011, bb:0010111101010, bb:1111110101, bb:11111110111110, bb:0001001110111, plc:nmod, bb:0011001111111, bb:1001001100, bb:1111011010, bb:0001101111001, bb:0001110000, bb:01101010, plh:sbj, w:n:time, bb:101111110, bb:00010100, pl:nmod, w:v:contact, pl:sbj	phw:in, bh:11011000, pl:pmod, bb:0001111010, bb:011011110, bb:1101101110, bb:01101001, bb:1111110100, pl:adv, bb:100111101110, bb:01110001, plh:adv, bb:11110100001111, bb:10001100, bb:0001101101010, bb:0001101111101, bb:11011000, bb:11110101110, plc:p, pcw:p
ENG:PROCRES	
bb:11111110011110, w:v:motion, bb:111100000100, pl:prd, plh:coord, w:v:social, bc:101101, p:nsiblings, bb:11111110101, bb:1111001101100, w:v:perception, bb:000011000, bb:01001100, bb:0000101100, bc:101100, bb:1111101100, bb:0001110000, bb:101011110, p:nchildren, m:plural	bc:11010, w:n:artifact, bb:110110110, bb:11010, pl:pmod, w:n:motive, w:n:possession, bb:010101001, bb:0010111010100, plh:vc, w:v:cognition, bb:11101101011, w:n:attribute, bb:1000111111, bb:000110111011, bb:100111101111, bb:00011100110, bb:001010100001, bb:0001000110, bb:00011001010

High positive coefficient	High negative coefficient
DA:CONTCONT	
bb:10001111110010, bb:110010100, bb:1100111111110110, bb:10001111110000, bb:1000111101010, bb:0111110, bb:100110100, bb:1010111110, bh:1101010100, pl:nobj, bb:1100101110, bb:1100100, bh:100110011, bb:10011101111, bh:11100011110, bb:1000110010111, bc:1000110010100, plh:dobj, bc:1000110010111, plc:nobj	phw:i, bh:01100, plh:root, bb:1101011110010, plc:pobj, bb:0000, bc:011101, bb:1110111100, bb:1110111010010, plc:rel, bb:1000110001110, w:container, bb:11101110000, bb:11010101011110, w:animal, bb:1101110101, w:languagerepresentation, bb:100000111111011, plc:coord, bb:1110011100
DA:LOCORG	
plh:subj, pl:dobj, bb:11101101110, bb:1101011001111, bb:1101011110110100, bb:100000111110000, bb:001011110, bb:10001111110100, bb:1110111110100, bb:11101111110111000, bb:1101110100, phw:med, bh:011101, bb:10000101100, w:group, bb:10000011100110, bb:1101011111000, bb:01101110, plh:vobj, pl:subj	pl:nobj, phw:i, bh:01100, plh:mod, plh:nobj, bb:101101000, bb:01111110, bb:10001110000, bb:10011110, p:rootdist, bb:11111110, bb:11011100, plh:lobj, bb:11011111, w:institution, bb:100011101110, bb:1100110, bb:11100110111, bb:1000101111000, bb:1000111101111
SPA:CONTCONT	
bb:1101111110111, bc:10010101111101, bb:1101010, bb:100111001111100, phw:beber, bb:1101111011110, bb:1001011111001, bb:10011010001, bc:11110, plh:do, w:n:cognition, bb:10011010000, w:v:consumption, plc:comp, bb:1100011011110010, bh:11000100111110, bb:10011010100, bb:1100010101101000, pcp:z, bb:100111111111000	bb:1011101111111100, phw:en, bh:1111100, pcp:d, bb:1100011111111101, bb:11000111010, bb:1100011011111100, bb:100111101110110, bb:10100111101, bb:1100011101111111, bc:0110, bb:10011111111011, bc:1011101111111100, bb:101100101100, bb:100110011110111, plh:mod, pcp:v, bb:10011010110110, bb:1100011011111111, w:n:tops
SPA:LOCORG	
pcp:d, bb:11111011110, bb:10111100010, bb:1011010011110, bb:101100111111110, bb:101111010, bb:10110001111, bb:1011010011010, w:v:competition, w:n:event, bb:100110110000, w:n:food, bb:100111101011010, plh:root, w:n:quantity, w:n:feeling, phw:con, bh:111111111, pl:subj, php:v	bh:1111100, phw:en, pl:mod, php:n, bb:1100000101110, bb:1101011110, bb:100001, bb:010, w:v:motion, bb:101011111110, bh:10101111011, bb:10101111000, bb:11000101010110, bb:10101111011, pcp:f, bb:1001110010101, bb:10011110010011, w:n:animal, bb:1001111000010, bb:1011001001

Appendix G

Agreement prediction

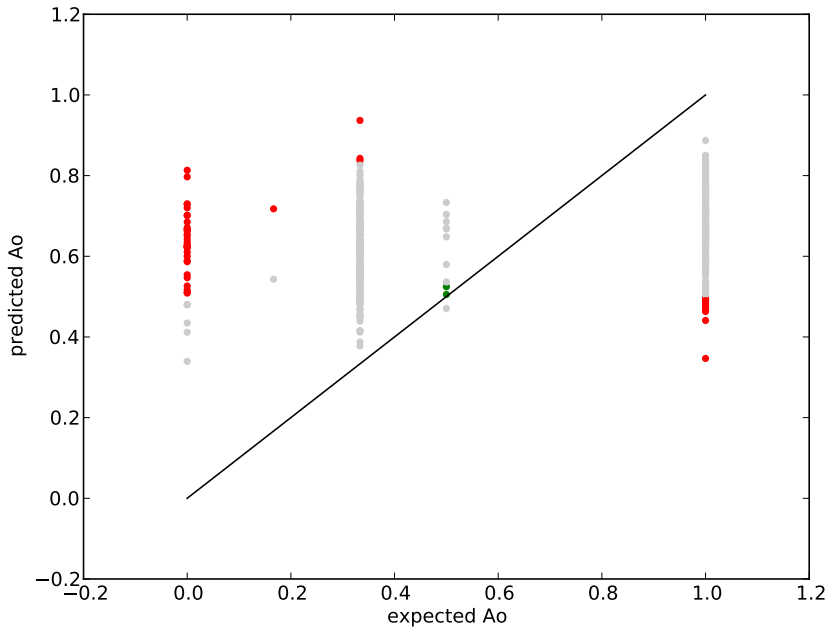

Figure G.1: Agreement prediction scatter plot for DK:CONTCONT

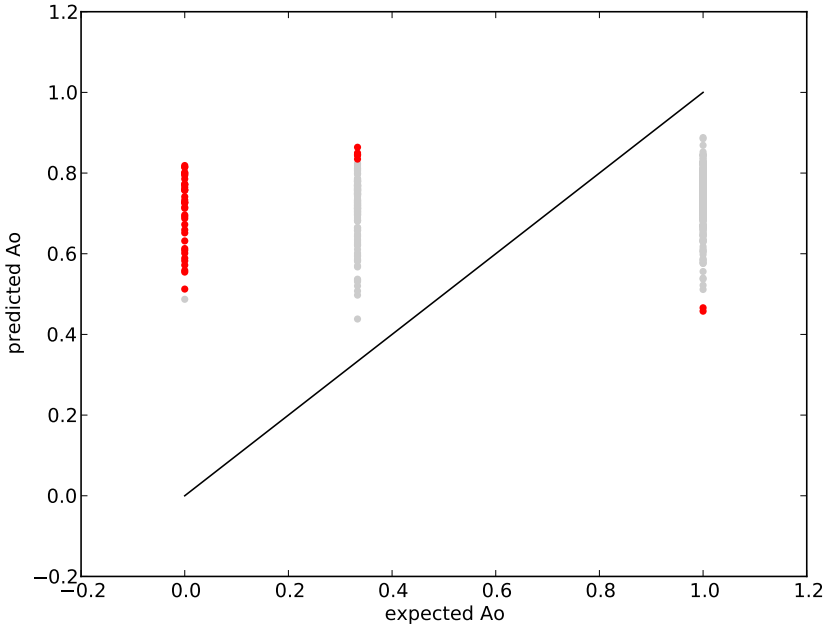

Figure G.2: Agreement prediction scatter plot for DK:LOCORG

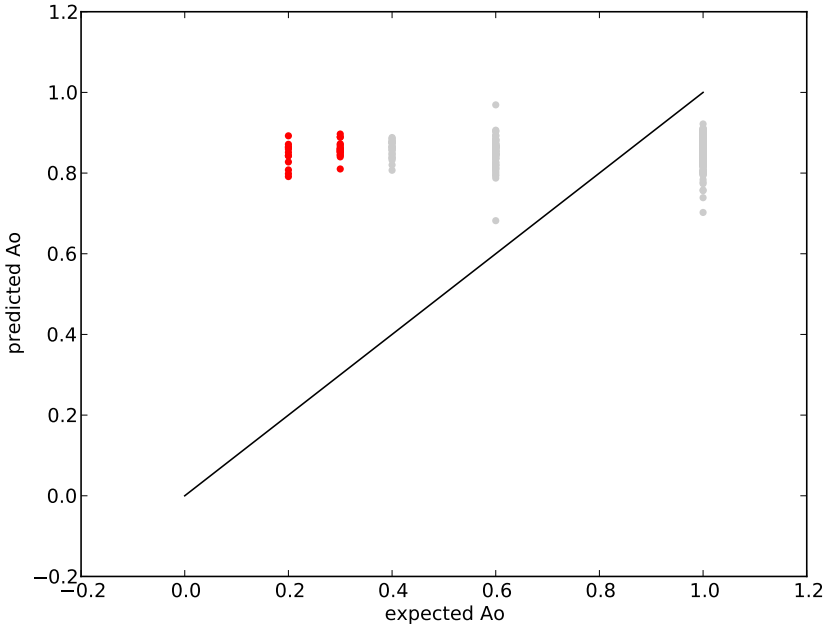

Figure G.3: Agreement prediction scatter plot for ENG:ANIMEAT

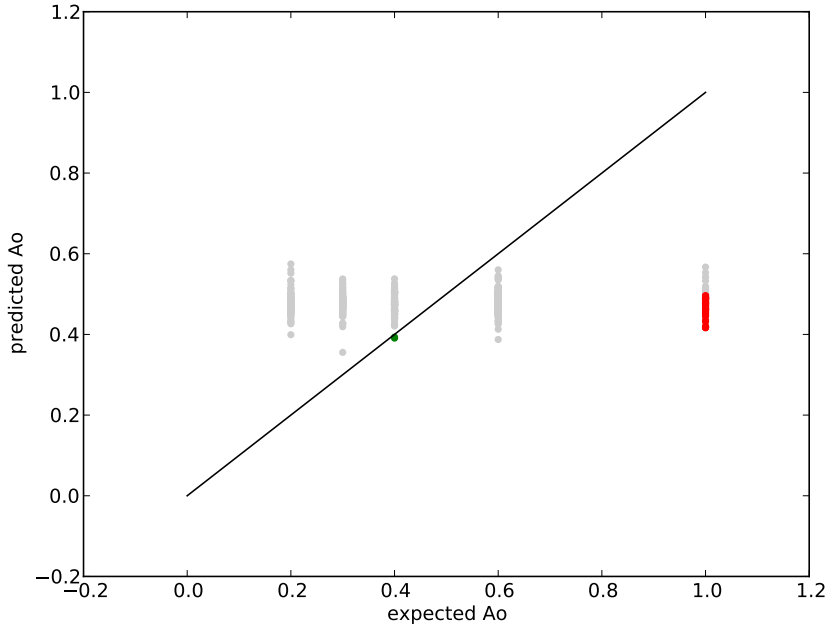

Figure G.4: Agreement prediction scatter plot for ENG:ARTINFO

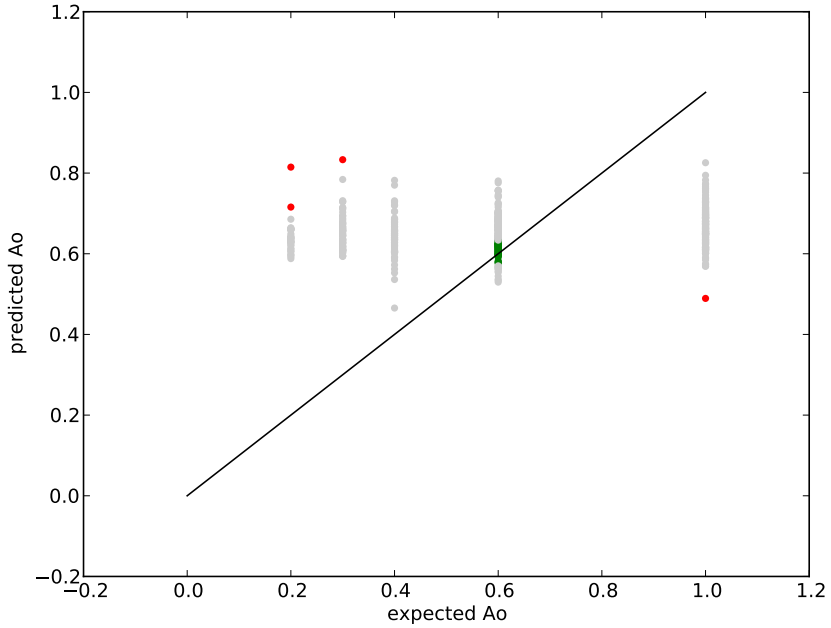

Figure G.5: Agreement prediction scatter plot for ENG:CONTCONT

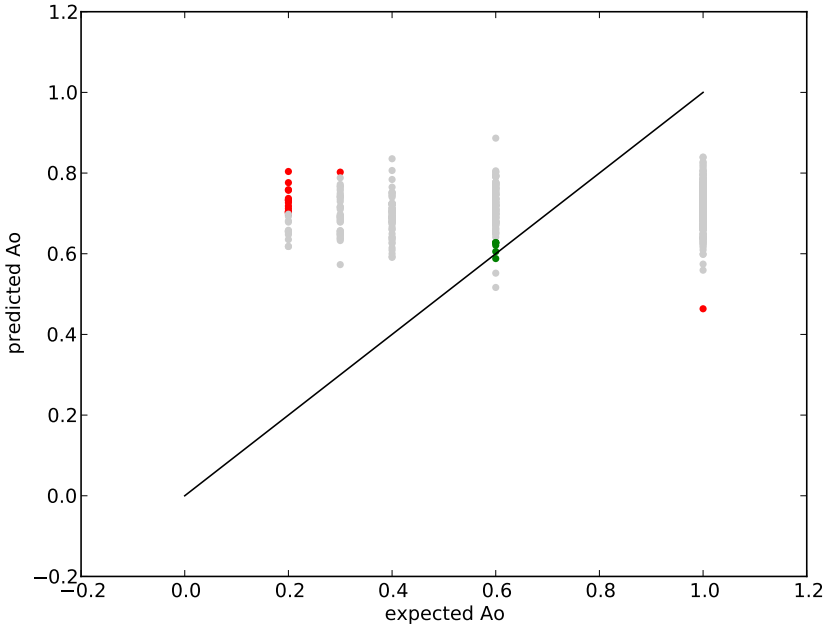

Figure G.6: Agreement prediction scatter plot for ENG:LOCORG

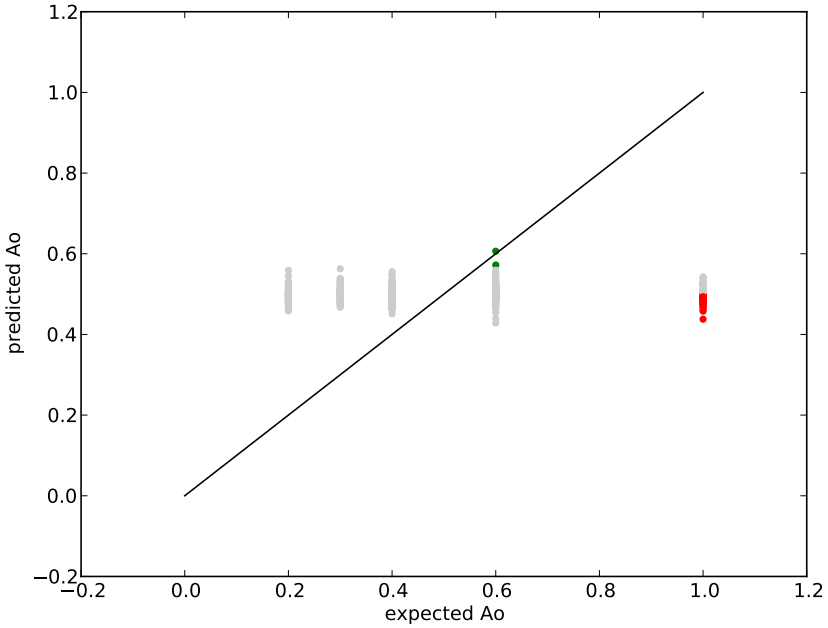

Figure G.7: Agreement prediction scatter plot for ENG:PROCRES

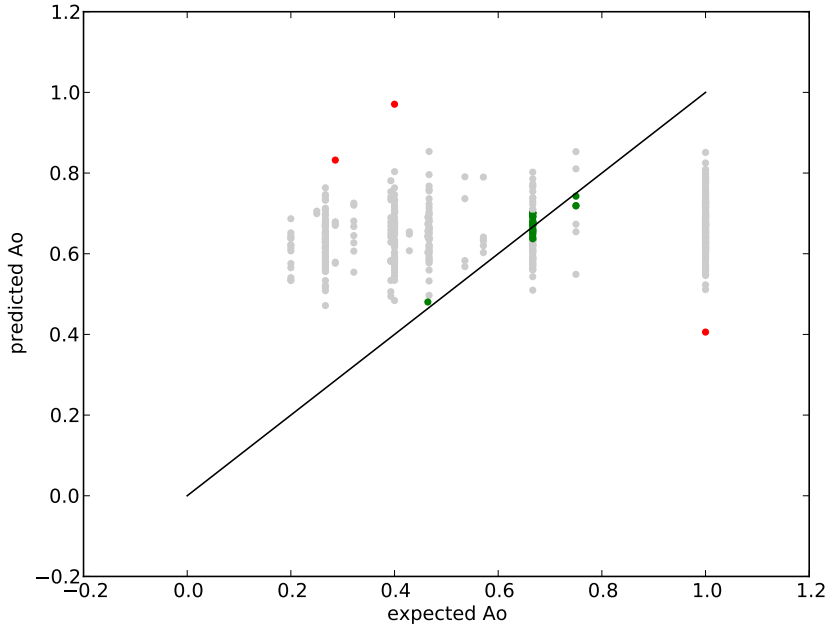

Figure G.8: Agreement prediction scatter plot for SPA:CONTCONT

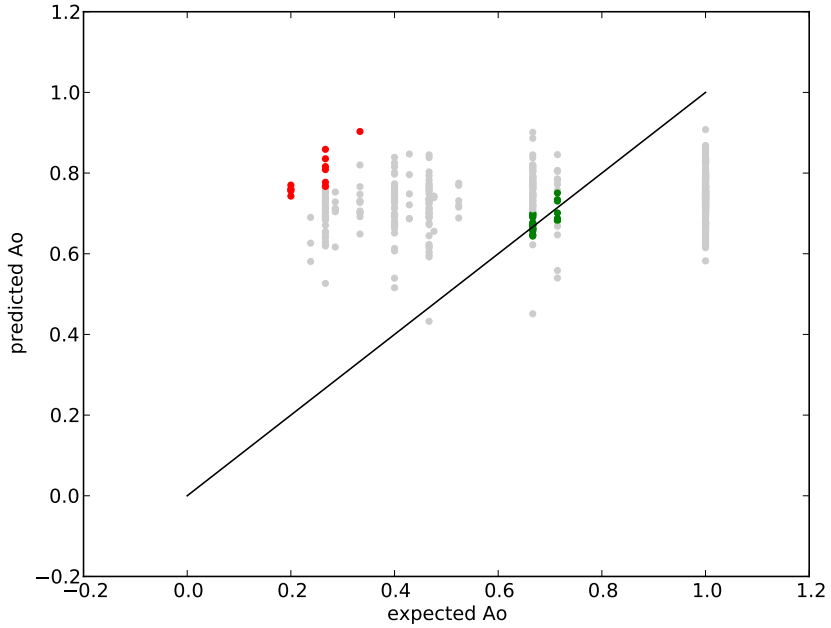

Figure G.9: Agreement prediction scatter plot for SPA:LOCORG

G.1 High-coefficient features for agreement

High positive coefficient	High negative coefficient
ENG:ANIMEAT	
bb:111100111101100, bb:0101101101, bb:0011000100, bb:11011000, bb:000100111010, bb:11011110, bb:1111000110100, bb:111111110, bb:11111101110, plh:sbj, bb:10011101110, w:v:cognition, w:v:consumption, pl:coord, bb:10001110, w:v:stative, w:v:creation, bb:0001110000, bb:1111100, w:n:location	bb:0000101100, m:plural, w:v:motion, w:n:event, p:nchildren, bb:1111111110, bb:01101010, bb:11011010, bb:10001010, w:n:act, bb:110110110, bb:0000100100, w:v:perception, bb:0001110001, plh:root, bb:00011101100, bb:101100, bb:1001110110, plh:coord, pl:sbj
ENG:ARTINFO	
bb:011100111, pl:obj, bb:10011101110, bb:0001001110110, bb:110110110, bb:1100010, bb:0111111, bb:11110010111110, bb:10001010, bb:011110, plc:nmod, bc:101100, w:n:feeling, bb:1111111110, w:n:act, w:v:creation, bb:00001011100, bb:11011000, w:n:motive, p:nchildren	bb:1111100, p:nsiblings, bc:101110, w:n:quantity, bb:111110100, w:n:event, w:n:cognition, bb:110000, w:n:body, bb:10011100, phw:and, bh:110000, bb:00001011111, w:v:cognition, w:n:time, bb:00011001110, bb:1111011010, bb:100111010, bb:00011100110, bb:00011101011
ENG:CONTCONT	
bb:11101111001, bb:01110001, bb:1101101110, bb:0101101101, bb:101100, bb:110110110, bc:101110, bb:11101111011, bc:11010, w:n:shape, bb:1111100, w:n:feeling, bb:111110111010, bb:0000100111, bb:11010, w:r:all, w:v:communication, plh:adv, bb:11011000, pl:pmod	w:n:cognition, bb:0001110000, p:nchildren, w:n:tops, w:v:perception, bb:0001110001, w:n:phenomenon, p:nsiblings, m:plural, bb:1001010111, bb:100011110, pl:coord, bb:101110, bb:00011101111, w:v:consumption, bb:11011111, w:n:plant, bb:11110101110, bc:10001110, bb:1111110101
ENG:LOCORG	
bb:11110100001101, bb:00010010111, bb:00010111110, bb:11111111110, bb:1001010111, bb:0001110011110, bb:00010110110, bb:1111101100, bb:0001110000, bb:10000010, w:n:plant, w:n:quantity, bb:000100111010, bh:11010, phw:of, bb:101110, w:n:time, plh:nmod, pl:nmod, w:n:motive	w:v:motion, plh:adv, bb:1101101110, bb:0, bb:11111101111, bb:1111100, bb:01110001, w:n:tops, p:nchildren, bb:011011110, bb:0001111010, bb:1100010, phw:in, bh:11011000, bb:0011110000, bb:0001101101100, w:n:group, bb:00010100, bb:10001010, pl:sbj
ENG:PROCRES	
bb:11111101111, bb:00011100100, bb:111100011110, bb:00000101110, w:v:motion, bb:110000, bb:001010010110, w:v:cognition, bb:0010111110110, bb:0010111011100, w:n:attribute, bb:11111011100, pl:pmod, pl:sbj, w:n:motive, w:n:food, w:n:event, bb:110110110, w:n:quantity, w:n:artifact	w:n:body, p:nsiblings, p:nchildren, bb:0001110000, bb:11011000, bb:101101, w:n:plant, bb:11111110101, pl:prd, w:n:communication, w:r:all, bb:11011110, bb:1100010, bb:00001010110, bb:11010, plc:p, pcw:p, bb:11011010, w:n:feeling, w:v:stative

High positive coefficient	High negative coefficient
DA:CONTCONT	
bc:1000110010111, w:comestible, bb:1101010100, bb:1001110110, p:nsiblings, bb:11111110, plh:vobj, bb:1100101100, bb:1110100110, bh:011101, phw:med, plh:dobj, bb:110010100, p:nchildren, w:possession, w:3rdorderentity, bb:011110, bb:1100101110, bc:100011001100, plc:nobj	bh:01100, phw:i, w:group, bb:01100, bb:1000110001010, bb:0111110, bb:110010101, bb:010, bb:011010, w:static, bb:100011000011, bb:1101011110 10, bb:0000, w:garment, bb:1110111010010, w:existence, w:substance, w:property, plh:coord, bb:11010111001001
DA:LOCORG	
bb:11001111101110, pl:subj, bb:10111010, plh:root, bb:11101101101, bb:1100100, bb:100010111110, w:group, w:agentive, bb:10110111111, bb:10000011100110, bb:1000001111110000, w:social, bb:101110011, w:geopoliticalplace, m:case:gen, w:mental, bb:10110111110, p:nsiblings, p:nchildren	bb:1011011101, bb:0111110, bb:11110, bb:01111110, bb:11011111, m:case:unmarked, plh:mod, w:unboundedevent, bb:1001110100, w:cause, bb:11000, bb:101110110, bb:10001111100, bb:10001010011, w:languagerepresentation, bb:10001110000, bb:1101010100, w:physical, bb:100011111101111, w:time
SPA:CONTCONT	
bb:110010, w:n:feeling, m:pos:ncmp000, bh:111111110, phw:por, bb:100000, w:n:location, p:nchildren, w:n:cognition, bb:1100011011101011, w:n:time, w:n:animal, plc:comp, bb:110110, bb:101111001110, phw:de, bh:11110, bc:11110, pcp:s, bb:10011010010	phw:en, bh:1111100, bb:101100101100, w:n:tops, w:a:all, bb:10101110, bb:1011101111111100, bb:10101111010, bb:1100011101111111, w:v:consumption, bb:111111110, pcp:a, bb:100110011110111, bb:11000100111110, w:v:contact, w:n:attribute, pcp:d, bc:010, w:v:change, bb:1100110
SPA:LOCORG	
bb:10011111000, bb:1001110011111110, bb:10010101000, bb:110111001011, bb:100111111111010, plh:comp, w:v:stative, pcp:n, bb:100110011100, w:a:pert, bb:11010010, bb:101110100, bb:00, bb:10011100000, w:n:event, w:v:competition, bh:11110, phw:de, bb:110110, w:a:all	bb:1111110, bb:0110, bb:1111100, w:n:body, w:v:possession, w:v:contact, w:n:object, bb:1001110010101, bb:11011101000, w:n:artifact, p:nsiblings, pl:mod, w:n:phenomenon, bb:11000110101100, w:n:cognition, bb:111110100, w:v:motion, bb:1001101011100, bb:1001110110, bb:1001111000100