Polysemy in verbs: systematic relations between senses and their effect on annotation

Anna Rumshisky
*Dept. of Computer Science
Brandeis University
Waltham, MA USA
arum@cs.brandeis.edu

Olga Batiukova†*
†Dept. of Spanish Philology
Madrid Autonomous University
Madrid, Spain
volha.batsiukova@uam.es

Abstract

Sense inventories for polysemous predicates are often comprised by a number of related senses. In this paper, we examine different types of relations within sense inventories and give a qualitative analysis of the effects they have on decisions made by the annotators and annotator error. We also discuss some common traps and pitfalls in design of sense inventories. We use the data set developed specifically for the task of annotating sense distinctions dependent predominantly on semantics of the arguments and only to a lesser extent on syntactic frame.

1 Introduction

Lexical ambiguity is pervasive in natural language, and its resolution has been used to improve performance of a number of natural language processing (NLP) applications, such as statistical machine translation (Chan et al., 2007; Carpuat and Wu, 2007), cross-language information retrieval and question answering (Resnik, 2006). Sense differentiation for the predicates depends on a number of factors, including syntactic frame, semantics of the arguments and adjuncts, contextual clues from the wider context, text domain identification, etc.

Preparing sense-tagged data for training and evaluation of word sense disambiguation (WSD) systems involves two stages: (1) creating a sense inventory and (2) applying it in annotation. Creating sense inventories for polysemous words is a task that is notoriously difficult to formalize. For polysemous verbs especially, constellations of related meanings make this task even more difficult. In lexicography, "lumping and splitting" senses during dictionary construction – i.e. deciding when to describe a set of usages as a separate sense – is a well-known problem (Hanks and Pustejovsky, 2005; Kilgarriff, 1997). It is often resolved on an ad-hoc basis, resulting in numerous cases of "overlapping senses", i.e. instances when the same occurrence may fall under more than one sense category simultaneously.

This problem has also been the subject of extensive study in lexical semantics, addressing questions such as when the context selects a distinct sense and when it merely modulates the meaning, what is the regular relationship between related senses, and what compositional processes are involved in sense selection (Pustejovsky, 1995; Cruse, 1995; Apresjan, 1973). A number of syntactic and semantic tests are traditionally applied for sense identification, such as examining synonym series, compatible syntactic environments, coordination tests such as *cross-understanding* or *zeugma* test (Cruse, 2000). None of these tests are conclusive and normally a combination of factors is used. At the recent Senseval competitions (Mihalcea et al., 2004; Snyder and Palmer, 2004; Preiss and Yarowsky, 2001), the choice of sense inventories frequently presented problems, spurring the efforts to create coarser-grained sense inventories (Hovy et al., 2006; Palmer et al., 2007; Navigli, 2006).

Part of the reason for such difficulties in establishing a set of senses available to a lexical item is that the meaning of a polysemous verb is often determined in composition and depends to the same extent on semantics of the particular arguments as it does on the base meaning of the verb itself. A number of systematic relations often holds between different senses of a polysemous verb. Depending on the kind of ambiguity involved in each case, some senses are easier to distinguish than others. Sense-tagged data (e.g. SemCor (Landes et al., 1998), PropBank (Palmer et al., 2005), OntoNotes (Hovy et al., 2006)) typically provides no way to differentiate between sense distinctions motivated by different factors. Treating different disambiguation factors separately would allow one to examine the contribution of each factor, as well as the success of a given algorithm in identifying the corresponding senses.

Within the scope of a sentence, syntactic frame and semantics of the arguments are most prominent in sense

Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, pages 33–41
Manchester, August 2008

disambiguation. The latter is often more subtle and hence complex. Our goal in the present study was to target sense distinctions motivated strongly or exclusively by differences in argument semantics. We base the present discussion on the sense-tagged data set we developed for 20 polysemous verbs. We argue below that cases which can not be reliably disambiguated by humans introduce noise into the data and therefore should be kept out, a principle adhered to in the design of this data set.

The choice of argument semantics as the target disambiguation factor was motivated by several considerations. In automatic sense detection systems, argument semantics is often represented using external resources such as thesauri or shallow ontologies. Sense induction systems using distributional information often do not take into account the possible implications of induced word clusters for sense disambiguation. Our goal was to analyze differences in argument semantics that contribute to disambiguation.

In this paper, we discuss different kinds of systematic relations observed between senses of polysemous predicates and examine the effects they have on decisions made by the annotators. We also examine sense inventories for other factors that influence inter-annotator agreement rates and lead to annotation error. In Section 2, we discuss some of the factors that influence compilation of sense inventories and the methodology involved. In Section 3, we describe briefly the data set and the annotation task. In Sections 4 and 5, we discuss the relations observed between different senses within sense inventories in our data set, their effect on decisions made by the annotators, and the related annotation errors.

2 Defining A Sense Inventory

Several current resource-oriented projects undertake to formalize the procedure of identifying a word sense. FrameNet (Ruppenhofer et al., 2006) attempts to organize lexical information in terms of script-like semantic frames, with semantic and syntactic combinatorial possibilities specified for each frame-evoking lexical unit (word/sense pairing). Semantics of the arguments is represented by Fillmore's case roles (*frame elements*) which are derived on ad-hoc basis for each frame.

In OntoNotes project, annotators use small-scale corpus analysis to create sense inventories derived by grouping together WordNet senses. The procedure is restricted to maintain 90% inter-annotator agreement (Hovy et al., 2006).

Corpus Pattern Analysis (CPA) (Hanks and Pustejovsky, 2005; Pustejovsky et al., 2004) attempts to catalog prototypical norms of usage for individual words, specifying them in terms of context patterns. As a corpus analysis technique, CPA has its origins in the analysis of large corpora for lexicographic purposes, of the kind that was used for compiling the Cobuild dictionary (Sinclair and Hanks, 1987). Each pattern gives a combination of surface textual clues and argument specifications. A lexicographer creates a set of patterns by sorting a concordance for the target predicate according to the context features. In the present study, we use a modification of the CPA technique in the way explained in Section 3.

In CPA, syntactic and textual clues include argument structure and minor syntactic categories such as locatives and adjuncts; collocates from wider context; subphrasal cues such as genitives, partitives, bare plural/determiner, infinitivals, negatives, etc. Semantics of the arguments is represented either through a set of shallow semantic types corresponding to basic semantic features (e.g. Person, Location, PhysObj, Abstract, Event, etc.) or extensionally through *lexical sets*, which are effectively collections of lexical items.[1]

Several CPA patterns may correspond to a single sense. The patterns vary in syntactic structure or the encoding of semantic roles relative to the described event. For example, for the verb *treat*, DOCTOR treating PATIENT and DOCTOR treating DISEASE both correspond to the medical sense of *treat*. Knowing which semantic role is expressed by a particular argument is often useful for performing inference. For instance, treating a disease eliminates the disease, but not the patient. In the present annotation task, each pattern is viewed as **sense in construction** and labeled as a separate sense. In the rest of the paper, we will use the term "sense" to refer also to such microsenses.

For the cases where sense differentiation depends strongly on differences in semantics of the arguments, several factors further complicate creating a sense inventory. Prototypicality as a general principle of category organization seems to play an important role in defining both the boundaries of senses and the corresponding argument groupings. The same sense of the predicate is often activated by a number of semantically diverse arguments. Such argument sets are frequently organized around a core of typical members that are a "good fit" with respect to semantic requirements of the corresponding sense of the target. The relevant semantic feature is prominent for them, while other, more peripheral members of the argument set, merely allow the relevant interpretation (see Rumshisky (2008) for discussion). For example, the verb *absorb* has a sense involving *absorbing a substance*, and the typical members of the corresponding argument set would be actual substances, such as *oil, oxygen, water, air, salt*, etc. But *goodness, dirt, flavor, moisture* would also activate the same sense.

Each decision to split a sense and make another category is to a certain extent an arbitrary decision. For example, for the verb *absorb*, one can separate *absorbing a substance* (*oil, oxygen, water, air, salt*) from *absorbing energy* (*radiation, heat, sound, energy*). The latter sense may or may not be separated from *absorb-*

[1] See Rumshisky et al. (2006) and Pustejovsky et al. (2004) for more detail.

ing impact (*blow, shock, stress*). But it is a marked continuum, i.e. certain points in the continuum are more prominent, with necessity of a given concept reflected in the frequency of use.

When several senses are postulated based on argument distinctions, there are almost always *boundary cases* that can be seen to belong to both categories. Consider, for example, two senses defined for the verb *launch* and the corresponding direct objects in (1):

(1) a. *Physically propel an object into the air or water*
missile, rocket, torpedo, satellite, shuttle, craft
b. *Begin or initiate an endeavor*
campaign, initiative, investigation, expedition, drive, competition, crusade, attack, assault, inquiry

The senses seem to be very clearly separated, yet examples like *launch a ship* clearly fall on the boundary: while *ships* are physical objects propelled into water, *launching a ship* can be virtually synonymous with *launching an expedition*.

Similarly, for the verb *conclude*, two senses below which are linked to nominal complements are clearly separated:

(2) a. *finish*
meeting, debate, investigation, visit, tour, discussion; letter, chapter, novel
b. *reach an agreement*
treaty, agreement, deal, contract, truce, alliance, ceasefire, sale

However, *conclude negotiations* is clearly a boundary case where both interpretations are equally possible (negotiations may be concluded without reaching an agreement). In fact, the two annotators chose different senses for this example:[2]

(3) We were able to operate under a lease agreement until purchase negotiations were concluded.
annoA: *finish*
annoB: *reach an agreement*

In many cases, postulating a separate sense for a coherent set of nominal complements is not justified, as there are regular semantic processes that allow the complements to satisfy selectional requirements of the verb. For example, the verb *conclude*, in the *finish* sense accepts EVENT complements. Therefore, nouns such as *letter, chapter, novel* in (2) must be coerced into events corresponding to the activity that typically brings them about, that is, re-interpreted as events of writing (their Agentive quale, cf. Pustejovsky (1995)). Similarly, the verb *deny* in the first sense (*state or maintain that something is untrue*) accepts PROPOSITION complements:

(4) a. *state or maintain that something is untrue*
allegations, reports, rumour; significance, importance, difference; attack, assault, involvement
b. *refuse to grant something*
access, visa, approval, funding, license

[2]All examples are taken from the annotated data set. In some cases, sentence structure was slightly modified for brevity.

Event nouns such as *attack* and *assault* are coerced into a propositional reading, as are relational nouns such as *significance* and *importance*.

Interestingly, as we have noted before (Rumshisky et al., 2006), each predicate imposes its own gradation with respect to prototypicality of elements of the argument set. As a result, even though basic semantic types such as PHYSOBJ, ANIMATE, EVENT, are used uniformly by many predicates, argument sets, while semantically similar, typically differ between predicates. For example, *fall* in the subject position and *cut* in the direct object position select for things that can be decreased:

(5) a. *cut (dobj)*: *reduce or lessen*
price, inflation, profits, cost, emission, spending, deficit, wages overhead, production, consumption, fees, staff
b. *fall (subj)*: *decrease*
price, inflation, profits, attendance, turnover, temperature, membership, import, demand, level

While there is a clear commonality between these argument sets, the overlap is only partial. To give another example, consider INFORMATION-selecting predicates *explain (subj)*, *grasp (dobj)* and *know (dobj)*. The nouns *book* and *note* occur in the subject position of *explain*; *answer* occurs both as the subject of *explain* and direct object of *know*; however, *grasp* accepts neither of these nouns as direct object. Thus, the actual selectional behavior of the predicates does not seem to be well described in terms of a fixed set of types, which is what is typically assumed by many ontologies used in automatic WSD.

3 Task Description

We were interested specifically in those cases where disambiguation needs to be made without relying on syntactic frame, and the main source of disambiguation is semantics of the arguments. Such cases are harder to identify formally in the development of sense inventories and harder for the annotators to determine. For example, phrasal verbs or idiomatic constructions that help identify a particular sense were intentionally excluded from our data set. Thus, for the verb *cut*, one of the senses involves cutting out a shape or a form (e.g. *cut a suit*), but the sentences with the corresponding phrasal form *cut out* were thrown out.

Even so, syntactic clues that contribute to disambiguation in some cases overrule the interpretation suggested by the argument. For example, for the verb *deny*, in *deny the attack*, the direct object strongly suggests a propositional interpretation for *deny* (that the attack didn't happen). However, the use of ditransitive construction (indicated in the example below by the past participle) overrules this interpretation, and we get the *refuse to grant* sense:

(6) Astorre, *denied* his *attack*, had stayed in camp, uneasily brooding.

In fact, during the actual annotation, one of the annotators did not recognize the use of past participle, and erroneously assigned the *state or maintain something to be untrue* sense to this sentence.

3.1 Data set

The data set was developed using the British National Corpus (BNC), which is more balanced than the more commonly annotated Wall Street Journal data. We selected 20 polysemous verbs with sense distinctions that were judged to depend for disambiguation on semantics of the argument in several argument positions, including direct object (dobj), subject (subj), or indirect object within a prepositional phrase governed by *with* (iobj_with):

dobj: *absorb, acquire, admit, assume, claim, conclude, cut, deny, dictate, drive, edit, enjoy, fire, grasp, know, launch*
subj: *explain, fall, lead*
iobj_with: *meet*

We used the Sketch Engine (Kilgarriff et al., 2004) both to select the verbs and to aid the creation of the sense inventories. The Sketch Engine is a lexicographic tool that lists collocates that co-occur with a given target word in the specified grammatical relation. The collocates are sorted by their association score with the target.

A set of senses was created for each verb using a modification of the CPA technique (Pustejovsky et al., 2004). A set of complements was examined in the Sketch Engine. If a clear division was observed between semantically different groups of collocates in a certain argument position, the verb was selected. For semantically distinct groups of collocates, a separate sense was added to the sense inventory for the target. For example, for the verb *acquire*, a separate sense was added for each of the following sets of direct objects:

(7) a. *Take on certain characteristics*
shape, meaning, color, form, dimension, reality, significance, identity, appearance, characteristic, flavor
b. *Purchase or become the owner of property*
land, stock, business, property, wealth, subsidiary, estate, stake

The sense inventory for each verb was cross-checked against several resources, including WordNet, PropBank, Merriam-Webster and Oxford English dictionaries, and existing correspondences in FrameNet (Ruppenhofer et al., 2006; Hiroaki, 2003), OntoNotes (Hovy et al., 2006),[3] and CPA patterns (Hanks and Pustejovsky, 2005; Rumshisky and Pustejovsky, 2006; Pustejovsky et al., 2004).

We performed test annotation on 100 instances, with the sense inventory additionally modified upon examining the results of the annotation. This sense inventory was provided to two annotators, along with 200 sentences for each verb. Each sentence was pre-parsed with RASP (Briscoe and Carroll, 2002), and the head of the target argument phrase was identified. Misparses were manually corrected in post-processing.

3.2 Defining the task for the annotators

Data set creation for a WSD task is notoriously hard (cf. Palmer et al. (2007)), as the annotators are frequently forced to perform disambiguation on sentences where no disambiguation can really be performed. This is the case, for example, for overlapping senses, where more than one sense is activated simultaneously (Rumshisky, 2008; Pustejovsky and Boguraev, 1993). The goal was to create, for each target word, a set of instances where humans had no trouble disambiguating between different senses.

Two undergraduate linguistics majors served as annotators. The annotators were instructed to mark each sentence with the most fitting sense. The annotators were allowed to mark the sentence as "N/A" and were instructed to do so if (i) the sense inventory was missing the relevant sense, (ii) more than one sense seemed to fit, or (iii) the sense was impossible to determine from the context.

With respect to metaphoric senses, instructions were to throw out cases of creative use where the interpretation was difficult or not immediately clear. The cases where the target grammatical relation was actually absent from the sentence also had to be marked as "N/A" (e.g. for *fire*, sentences without direct object, e.g. *a stolen car was fired upon*). The annotators were also instructed to mark idiomatic expressions and phrasal verbs as "N/A", e.g. for the verb *fall*: *fall from favor, fall through, fall in, fall back, fall silent, fall short, fall in love*.

Disagreements between the annotators were resolved in adjudication by the co-authors. The average inter-annotator agreement (ITA) for our data set was computed as a macro-average of the percentage of instances that were annotated with the same sense by both annotators to the total number of instances retained in the data set for each verb. The instances that were marked as "N/A" by one of the annotators (or thrown out during the adjudication) were not included in the computation. The ITA value for our data set was 95%. However, as we will see below, the ITA values do not always reflect the actual accuracy of annotation, due to some common problems with sense inventories.

3.3 Glossing a sense

A very common problem with glossing a sense involves the situation where a sense inventory includes two senses one of which is an extension of the other. The derived sense may be related to the primary sense through metaphor, and this often results in the former taking on a semantically less specific interpretation. The problem with creating glosses in this situation is that the words used may have sense distinctions

[3] Sense inventories released for the 65 verbs made available for SemEval-2007.

parallel to the ones in the target verb being described. This leaves the annotators free to choose either sense. This seems to be the case, for example, with OntoNotes sense inventory for *fire*, where *ignite or become ignited* is the gloss under which very divergent examples are grouped: *oil fired the furnace* (literal, primary sense) and *curiosity fired my imagination* (metaphoric extension). Clearly, annotators were having a problem with this sense due to the fact that the verb *ignite* has sense distinctions which are based on the same metaphor (*fire* = *inspire*) and therefore are very similar to those of the verb *fire*.

In case of semantic underspecification, annotators may be left free to choose the more generic sense, which contaminates the data set while not being reflected in the inter-annotator agreement values. For example, in our sense inventory for *acquire*, the gloss for *acquire a new customer* has to be very generic. We used the gloss "become associated with something, often newly brought into being". However, that led the annotators to overuse this gloss and select this sense in cases where a more specific gloss was more appropriate:[4]

(8) By this treaty, Russia *acquired* a Black Sea *coastline*.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *purchase or become the owner of property*

For a more detailed analysis of this phenomenon, see Section 5.

4 Relations Between Senses

In this section, we discuss linguistic processes underlying relations between senses within a single sense inventory. We believe that a detailed analysis of these processes should help to account for the annotator's ability to perform disambiguation. Some sense distinctions appear more striking to the annotators, depending on the type of relation involved.

In line with existing approaches to sense relations, we will look at both the linguistic structures involved in sense modification and the productive processes acting on linguistic structures. For the purposes of our present discussion, we interpret the literal (physical, direct) senses to be primary, with respect to more abstract or metaphorical senses.

4.1 Argument structure alternations

Some of the most striking differences between the senses are related to the argument structure alternations:

1. Different case roles (frame elements) may be expressed in the same argument position (in this case, direct object), corresponding to different perspectives on the same event. For example, direct object position of the verb *drive* may be filled by VEHICLE, DISTANCE, or PHYSOBJ giving rise to three distinct senses: (i) *operate a vehicle controlling its motion*, (ii) *travel in a vehicle a certain distance*, and (iii) *transport something or someone*. Similarly, for the verb *fire*, PROJECTILE or WEAPON in direct object position give rise to two related senses: (i) *shoot, discharge a weapon*, (ii) *shoot, propel a projectile*.

2. The distinction between propositional and non-propositional complements, as for the verbs *admit* and *deny* in (9) and (10):

(9) a. *admit defeat, inconsistency, offense*
(*acknowledge the truth or reality of*)
b. *admit patients, students*
(*grant entry or allow into a community*)

(10) a. *deny reports, importance, allegations*
(*state or maintain to be untrue*)
b. *deny visa, access*
(*refuse to grant*)

3. There is a mutual dependency between subcategorization features of the complements in different argument positions. For example, the [+animate] subject may combine with specific complements not available for [−animate], as for the two senses of *acquire*: (i) *learn* and (ii) *take on certain characteristics*. Compare NP_{subj} [-animate] *acquire* NP_{dobj} (*language, manners, knowledge, skill*) vs. NP_{subj} [−animate] *acquire* NP_{dobj} (*importance, significance*). Similarly, for *absorb*, compare NP_{subj} [±animate] *absorb* NP_{dobj} (*substance*) and NP_{subj} [+animate] *absorb* NP_{dobj} (*skill, information*). Note that, as one would expect, such dependencies are inevitable even despite the fact that our data set was developed specifically to target sense distinctions dependent on a single argument position.

4.2 Event structure modification

Event structure modifications (i.e. operations affecting aspectual properties of the predicate) are another source of sense differentiation. Two cases appear most prominent:

1. The event structure is modified along with the characteristics of the arguments. For example, for *enjoy*, compare *enjoy skiing, vacation* (DYNAMIC EVENT) with *enjoying a status* (STATE). Similarly, for *lead*, compare *a person leads smb somewhere* (PROCESS) vs. *a road* (PATH) *leads somewhere* (STATE); for *explain*, compare *something or somebody explains smth* (= *clarifies, describes, makes comprehensible*, PROCESS) vs. *something* [−inanimate, +abstract] *explains something* (= *is a reason for something*, STATE); for *fall*, compare PHYSOBJ *falls* (TRANSITION or ACCOMPLISHMENT) vs. *a case falls into a certain category* (STATE).

2. The aspectual nature of the predicate is the only semantically relevant feature that remains unchanged after consecutive sense modifications. For example, the ingressive meaning of 'beginning something' is preserved in shifting from the physical sense of the verb *launch* in *launch a missile* to *launch a campaign* and *launch a product*.

[4] We will refer to annotators A and B as *annoA* and *annoB*.

4.3 Lexical semantic features

Sense distinctions often involve deeper semantic characteristics of the verbs which could be accounted for by means of lexical semantic features such as qualia structure roles in Generative Lexicon (Pustejovsky, 1995):[5]

1. Consider how the meaning component 'manner of motion' (typically associated with the agentive role) gets transformed in the different senses of *drive*. It is obviously present in the physical uses of *drive* (such as *operate a vehicle, transport something or somebody*, etc.), but is completely lost in *motivate the progress of* (as in *drive the economy, drive the market forward*, etc.). The value of the agentive role of *drive* becomes underspecified or semantically weak, so that the overall meaning of *drive* is transformed to *cause something to move.*

2. Information about semantic type contained in qualia structure allows apparently diverse elements to activate the same sense of the verb. For instance, the verb *absorb* in the sense *learn or incorporate skill or information* occurs with direct objects such as *values, atmosphere, information, idea, words, lesson, attitudes, culture*. The requisite semantic component is realized differently for each of these words. Some of them are complex types[6] with INFORMATION as one of the constituent types: *words* (ACOUSTIC/VISUAL ENTITY • INFO), *lesson* (EVENT • INFO). Others, such as *idea*, are polysemous, with one of the senses being INFORMATION. Cases like *culture* and *values* are more difficult, but since they refer to knowledge, the INFORMATION component is clearly present. Consequently, the annotators are able to identify the corresponding sense of *absorb* with a high degree of agreement.

4.4 Metaphor and metonymy

The processes causing the mentioned meaning transformations in our corpus often involve metaphor and metonymy. Below are some of the conventionalized extensions with metaphorical flavor:

(11) a. *grasp object* vs. *grasp meaning*
b. *launch object* vs. *launch an event (campaign, assault)* or *launch a product (newspaper, collection)*
c. *meet with a person* vs. *meet with success, resistance*
d. *lead somebody somewhere* vs. *lead to a consequence*

Note that these metaphorical extensions involve abstract or continuous objects (*meaning, assault, success, consequence*), which in turn cause event structure modifications (*lead* as a process vs. *lead* as a state). Thus, the processes and structures we are dealing with are clearly interrelated.

The metonymical process can be exemplified by *edit* as *make changes to the text* and as *supervise publication*, which are in a clear contiguity relationship.

One of the effects of the metaphorization and progressive emptying of the primary (physical, concrete) senses is the distinction between generic and specific senses. For example, compare *acquire land, business* (specific sense) to *acquire an infection, a boyfriend, a following*, which refers to some extremely light generic association. Similar process is observed for the semantically weak sense of *fall, be associated with or get assigned to a person or location or for event to fall onto a time*:

(12) Birthdays, lunches, celebrations *fall* on a certain date or time
Stress or emphasis *fall* on a given topic or a syllable
Responsibility, luck, suspicion *fall* on or to a person

The specificity often involves specialization within a certain domain:

(13) a. *conclude* as *finish* vs. *conclude* as *reach an agreement* (Law, Politics)
b. *fire* as *shoot a weapon or a projectile* vs. *fire* as *kick or pass an object of play in sports* (Sport)

Thus, when concluding a *pact* or an *agreement*, a certain EVENT is also being finished (negotiation of that agreement), necessarily with a positive outcome.

In the following section, we will try to show how different kinds of relations between senses influence disambiguation carried out by the annotators. In particular, we look at different sources of disagreement and annotator error as determined in adjudication.

5 Analysis of Annotation Decisions

As we have seen above, in many cases disambiguation is impossible due to the nature of compositionality. Also, as there are no clear answers to a number of questions concerning sense identification, the annotators deal with sense inventories that are imperfect. Results of the disambiguation task carried out by the annotators reflect all these defects.

In cases when a specific meaning from the data set is not included into the sense inventory (e.g. due to its low frequency or extreme fine-grainedness) the annotators may use a more general meaning or pick the closest meaning available. For example, within the sense inventory for *fire*, there was no separate gloss for *fire an engine.* Annotator A in our experiment chose the closest specific meaning available, and Annotator B marked it with a more generic sense:

(14) Engineers successfully *fired* thrusters to boost the research satellite to an altitude of 507 km.
annoA: *shoot, propel a projectile*
annoB: *apply fire to*

As mentioned in Section 3.3, even when the appropriate specific sense is available, annotators frequently chose the more generic sense in its place, as in (15), (16) and (17), and also in (8).

[5] We will use the terminology from Generative Lexicon (Pustejovsky, 1995; Pustejovsky, 2007) to discuss lexical semantic properties, such as *qualia roles*, *complex* and *functional types*, and so on.

[6] *Complex type* is a term used for concepts that inherently refer to more than one semantic type.

(15) Several *referrals fell* into this *category*.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *be categorized as or fall into a range*

(16) The terrible *silence* had *fallen*.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *for a state (such as darkness or silence) to come, to commence*

(17) He *acquired* a *taste* for performing in public.
annoA: *become associated with something, often newly brought into being*
annoB: *become associated with something, ...*
correct: *learn*

Note that in (8) this decision was probably motivated by the annotators' uncertainty about the semantic ascription of the relevant argument (*coastline* is not a prototypical owned property). The generic sense seems to be the safest option to take for the annotators, as compared to taking a chance with a specific meaning. Due to its low degree of semantic specification, the generic sense is potentially able to embrace almost every possible use. This is not a desirable outcome because the generic senses are introduced in the inventory to account only for semantically underspecified cases. For instance, *become associated with something, often newly brought into being* is appropriate for *acquire a grandchild*, but not for *acquire a taste* or *acquire a proficiency*.

Remarkable variation is also observed with respect to **non-literal uses** as discussed in Section 4.4. For example, in (18) and (19) abstract NPs *panic* and *imbalance of forces* are equated with *energy or impact* by one annotator and with *substance* by the other.

(18) Her *panic* was *absorbed* by his warmth.
annoA: *absorb energy or impact*
annoB: *absorb substance*

(19) Alternatively, *imbalance* of forces can be *absorbed* into the body.
annoA: *absorb energy or impact*
annoB: *absorb substance*

In some cases, the literal and the metaphoric senses are activated simultaneously resulting in ambiguity (cf. Cruse (2000)):

(20) For over 300 years this waterfall has provided the energy to *drive* the *wheels* of industry.
annoA: *motivate the progress of*
annoB: *provide power for or physically move a mechanism*

(21) But fashion changed and the short *skirt fell* – literally – from favour and started skimming the ankles.
annoA: *lose power or suffer a defeat*
annoB: N/A

(22) She was delighted when the *story* of Hank *fell* into her lap.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *physically drop; move or extend downward*

Impact of **subcategorization features** on disambiguation (cf. Section 4.1 para 3) is illustrated in (23).

(23) The reggae tourist can easily *absorb* the current reggae *vibe*.
annoA: *absorb energy or impact*
annoB: *learn or incorporate skill or information*

Both interpretations chosen here (*absorb energy or impact* and *learn or incorporate skill or information*) were possible due to the animacy of the subject, which activates two different subcategorization frames and subsequently two different senses.

Typically, cases where **semantic type** of the relevant arguments (cf. Section 4.3 para 2) is not clear result in annotator disagreement:

(24) The AAA *launched* education *programs*.
annoA: *begin or initiate an endeavor* (EVENT)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

(25) France plans to *launch* a remote-sensing *vehicle* called Spot.
annoA: *physically propel into the air, water or space* (PHYSOBJ)
annoB: *begin to produce or distribute; start a company* (PRODUCT)

The two cases above are interesting in that both *program* and *vehicle* are ambiguous and can be analyzed semantically as members of different semantic classes. This is what the annotators in fact do, and as a result, ascribe them to different senses. *Program* can be categorized as EVENT ('series of steps') or as INTELLECTUAL ACTIVITY PRODUCT ('document or system of projects'). It is a complex type, i.e. it is an inherently polysemous word that represents at least two different semantic types. *Vehicle*, in turn, is a functional type: on the one hand, it represents an entity with certain formal properties (PHYSOBJ interpretation), on the other hand, it is an artifact, with a prominent practical purpose (PRODUCT interpretation).

In fact, most problems the annotators had with the task are due to the inherent semantic complexity of words such as *vehicle* and *program* in (24) and (25) and to the existence of boundary cases, where the relevant noun does not properly belong to one or another semantic category. This is the case with *panic*, *imbalance* or *reggae vibe* in (18), (19), and (23), and also with *taste* and *coastline* in (17) and (7).

In some of these cases, other contextual clues may come into play and tip the balance in favor of one or another sense. Note that disambiguation was influenced by a **wider context** even despite the intentionally restrictive task design (targeting a particular syntactic relation for each verb). For instance, in (26), **domain-specific clues** referring to war or military conflict (such as *rebel control*) could have motivated Annotator B's decision to ascribe it to the sense *lose power or suffer a defeat* (even though a road is not typically an entity that can lose power), while the other annotator chose a more generic meaning:

(26) The *road fell* into rebel control.
annoA: *be associated with or get assigned to a person or location or for event to fall onto a time*
annoB: *lose power or suffer a defeat*

Other pragmatic and discourse-oriented clues played a role, in particular, positive and negative connotation of the senses and the relevant arguments, as well as the temporal organization of discourse. For example, in (27) and (28), positive or neutral interpretation of *wave of immigrants* and *change* could have led to the choice of *take in or assimilate* and *learn or incorporate skill or information* senses, while the negatively-colored interpretation might explain the choice of the *bear the cost of* sense.

(27) ..help *absorb* the latest *wave of immigrants*.
annoA: *bear the cost of; take on an expense*
annoB: *take in or assimilate, making part of a whole or a group*

(28) For senior management an important lesson was the trade unions' capacity to *absorb change* and to become its agents.
annoA: *learn or incorporate skill or information*
annoB: *bear the cost of; take on an expense*

Temporal organization of a broader discourse is another important factor. For example, for the verb *claim*, the senses *claim the truth of* and *claim property you are entitled to* have different presuppositions with respect to preexistence of the thing claimed. In (28), due to the absence of a broader context, the annotators chose two different temporal reference interpretations. For Annotator B, *success* was something that has happened already, while for A this was not clear (*success* might have been achieved or not):

(29) One area where the government can *claim* some *success* involves debt repayment.
annoA: *come in possession of or claim property you are entitled to*
annoB: *claim the truth of*

6 Conclusion

We have given a brief overview of different types of sense relations commonly found in polysemous predicates and analyzed their effect on different aspects of the annotation task, including sense inventory design and execution of the WSD annotation.

The present analysis suggests that theoretical tools must be refined and further developed in order to give an adequate account to the sense modifications found in real corpus data. To this end, broader contextual clues and discourse-oriented clues need to be included in the analysis.

Semantically annotated corpora are routinely developed for the training and testing of automatic sense detection and induction algorithms. But they do not typically provide a way to distinguish between different kinds of ambiguities. Consequently, it is difficult to perform adequate error analysis for different sense detection systems. Appropriate semantic annotation that would allow one to determine which sense distinctions can be detected better by automatic systems does not need to be highly specific and unnecessarily complex, but requires development of robust generalizations about sense relations.

One obvious conclusion is that data sets need to be explicitly restricted to the instances where humans have no trouble disambiguating between different senses. Thus, prototypical cases can be accounted for reliably, ensuring the clarity of annotated sense distinctions. At face value, imposing such restrictions may appear to negatively influence usability of the resulting data set in particular applications requiring WSD, such as machine translation or information retrieval. However, this decision impacts most strongly those boundary cases which are not reliably disambiguated by human annotators, and which rather introduce noise into the data set.

Acknowledgments

This work was supported in part by NSF CRI grant to Brandeis University. The work of O. Batiukova is supported by postdoctoral grant of the Ministry of Education of Spain and Madrid Autonomous University.

References

Apresjan, Ju. 1973. Regular polysemy. *Linguistics*, 142(5):5–32.

Briscoe, T. and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, May 2002*, pages 1499–1504.

Carpuat, M. and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of EMNLP-CoNLL*, pages 61–72.

Chan, Y. S., H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*, pages 33–40, Prague, Czech Republic, June.

Cruse, D. A. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Dizier, Patrick St. and Evelyne Viegas, editors, *Computational Lexical Semantics*, pages 33–49. Cambridge University Press, Cambridge, England.

Cruse, D. A. 2000. *Meaning in Language, an Introduction to Semantics and Pragmatics*. Oxford University Press, Oxford, United Kingdom.

Hanks, P. and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*.

Hiroaki, S. 2003. FrameSQL: A software tool for FrameNet. In *Proceedigns of ASIALEX '03*, pages 251–258, Tokyo, Japan. Asian Association of Lexicography.

Hovy, E., M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Kilgarriff, A., P. Rychly, P. Smrz, and D. Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex, Lorient, France*, pages 105–116.

Kilgarriff, A. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

Landes, S., C. Leacock, and R.I. Tengi. 1998. Building semantic concordances. In Fellbaum, C., editor, *Wordnet: an electronic lexical database*. MIT Press, Cambridge (Mass.).

Mihalcea, R., T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In Mihalcea, Rada and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July. Association for Computational Linguistics.

Navigli, R. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia, July. Association for Computational Linguistics.

Palmer, M., D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Palmer, M., H. Dang, and C. Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Journal of Natural Language Engineering*.

Preiss, J and D. Yarowsky, editors. 2001. *Proceedings of the Second Int. Workshop on Evaluating WSD Systems (Senseval 2)*. ACL2002/EACL2001.

Pustejovsky, J. and B. Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artif. Intell.*, 63(1-2):193–223.

Pustejovsky, J., P. Hanks, and A. Rumshisky. 2004. Automated Induction of Sense in Context. In *COLING 2004, Geneva, Switzerland*, pages 924–931.

Pustejovsky, J. 1995. *Generative Lexicon*. Cambridge (Mass.): MIT Press.

Pustejovsky, J. 2007. Type Theory and Lexical Decomposition. In Bouillon, P. and C. Lee, editors, *Trends in Generative Lexicon Theory*. Kluwer Publishers (in press).

Resnik, P. 2006. Word sense disambiguation in NLP applications. In Agirre, E. and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Rumshisky, A. and J. Pustejovsky. 2006. Inducing sense-discriminating context patterns from sense-tagged corpora. In *LREC 2006, Genoa, Italy*.

Rumshisky, A., P. Hanks, C. Havasi, and J. Pustejovsky. 2006. Constructing a corpus-based ontology using model bias. In *The 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida, USA.

Rumshisky, A. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. *Distributional Models of the Lexicon in Linguistics and Cognitive Science, special issue of Italian Journal of Linguistics / Rivista di Linguistica*. forthcoming.

Ruppenhofer, J., M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

Sinclair, J. and P. Hanks. 1987. *The Collins Cobuild English Language Dictionary*. HarperCollins, 4th edition (2003) edition. Published as Collins Cobuild Advanced Learner's English Dictionary.

Snyder, B. and M. Palmer. 2004. The english all-words task. In Mihalcea, Rada and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.