Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1680–1690
Marseille, 20-25 June 2022

Polysemy in Spoken Conversations and Written Texts

Aina Garí Soler, Matthieu Labeau, Chloé Clavel
LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.garisoler,matthieu.labeau,chloe.clavel}@telecom-paris.fr

Abstract

Our discourses are full of potential lexical ambiguities, due in part to the pervasive use of words having multiple senses. Sometimes, one word may even be used in more than one sense throughout a text. But, to what extent is this true for different kinds of texts? Does the use of polysemous words change when a discourse involves two people, or when speakers have time to plan what to say? We investigate these questions by comparing the polysemy level of texts of different nature, with a focus on spontaneous spoken dialogs; unlike previous work which examines solely scripted, written, monolog-like data. We compare multiple metrics that presuppose different conceptualizations of text polysemy, i.e., they consider the observed or the potential number of senses of words, or their sense distribution in a discourse. We show that the polysemy level of texts varies greatly depending on the kind of text considered, with dialog and spoken discourses having generally a higher polysemy level than written monologs. Additionally, our results emphasize the need for relaxing the popular "one sense per discourse" hypothesis.

Keywords: Semantics, Word Sense Disambiguation, Document classification / Text categorisation

1. Introduction

Polysemous words are not the majority in the vocabulary of a language, but they are used frequently in all types of texts. These words may even be employed in their different senses within the same discourse without necessarily hindering communication. For example, speakers sometimes exploit polysemy to keep a conversation interesting, with irony, jokes, or metaphors (Nerlich and Clarke, 2001).

Gale et al. (1992) investigate whether multiple instances of a polysemous word in a discourse tend to share the same sense, and conclude that this is indeed the case. They propose the "one sense per discourse" (OSD) hypothesis, which, despite its popularity and convenience to be used as a heuristic for Word Sense Disambiguation (McCarthy and Carroll, 2003; Preiss and Stevenson, 2013; Pilehvar and Navigli, 2015; Chaplot and Salakhutdinov, 2018), has also been put into question (Krovetz, 1998).

Works exploring the use of polysemous words in texts (Magnini et al., 2002; McCarthy et al., 2007) have focused on a specific kind of discourse, namely scripted written monologs. In this paper, we want to compare the polysemy level of different kinds of texts. We hypothesize that the use of polysemous words in multiple senses within a discourse varies depending on the nature of the text. In order to verify this hypothesis, we compare the polysemy level of texts along three different axes: spoken/written, monolog/(dyadic) dialog, and spontaneous/scripted. In our analysis, the most prominent type of discourse is spoken language transcripts, and most of these consist of dialogs. There are reasons to expect the polysemy level of this type of discourses to be higher, or at least different, than that of its counterpart, i.e., monologs. Differences between speakers' background, world knowledge, idiolect, language level, or even opinions may pose challenges to communication (Pronin et al., 2002; Kaur, 2011). Speakers need to make an effort to reach mutual understanding by, for example, adapting to their interlocutor and backchanneling (Clark and Brennan, 1991; Pickering and Garrod, 2006; Liberman, 2012). This, along with possible misunderstandings (Bazzanella and Damiano, 1999), disagreements or even jokes, may result in words being used in multiple senses. At the same time one could also argue that, in monologs, precisely due to the lack of feedback, the speaker may not be conscious of ambiguities in their discourse. Dialog participants, instead, deliberately try to avoid ambiguity to facilitate understanding, by restricting recurring expressions to a preestablished meaning (Garrod and Anderson, 1987).

Another distinction that we make is between spontaneous and scripted language. We expect scripted texts to display a lower level of polysemy than spontaneous ones, because there is more time to carefully plan the use of words to avoid unnecessary repetitions and ambiguity. Improvised texts would be more likely to contain disfluencies and, in the case of dialogs, overlapped speech (Busso and Narayanan, 2008).

Knowing the polysemy level of a given discourse set can be useful not only to better characterize it and to compare discourse sets of different nature, but also to adapt disambiguation strategies to the properties of a given text. When it comes to conversations, it allows to investigate, for the first time, the relationship between polysemy and other factors like dialog success (Friedberg et al., 2012) or interpersonal rapport (Sinha and Cassell, 2015).

Our methodology relies on existing measures of polysemy, as well as on a new measure that we propose. Importantly, we distinguish between measures of *observed* polysemy (the focus of this paper), which take into account the senses in which words are used in a text; from *potential* polysemy, which simply deter-

mines the polysemy level of words in a text based on an external resource such as WordNet (Fellbaum, 1998). We compare the polysemy level of different types of texts, from written monologs to spontaneous spoken conversations. We use WordNet synsets and also experiment with coarser-grained senses.
Our findings confirm our initial hypothesis that different kinds of texts exhibit different polysemy levels, and provide more evidence against the widely adopted OSD idea. Specifically, dialogs and spoken discourses tend to have higher polysemy than written texts. The main contributions of this paper are as follows:

- For the first time, we compare the observed polysemy level of sets of texts of different nature;
- We compare multiple existing textual polysemy measures and propose our own;
- We investigate the relation between polysemy and task success in task-oriented conversations.

2. Background

2.1. Textual Polysemy

Different measures to calculate textual polysemy have been proposed. We can distinguish between measures of what we refer to as *observed* and *potential* polysemy. The former is concerned with the actual diversity of senses used in a text. With potential polysemy we instead refer to the a priori ambiguity of words in a text. These two notions of polysemy offer two related, but distinct, perspectives. Potential polysemy provides an indication of the difficulty of performing automatic WSD on a text, while observed polysemy determines how many of these senses are actually used within a discourse.
The most common measure of potential polysemy consists in calculating the average number of senses that words in a text can have according to a sense inventory (Graesser et al., 2004; Pasini and Camacho-Collados, 2020). The proportion of monosemous or polysemous words can also be used (Agirre and Rigau, 1996). Potential polysemy has been more extensively studied than observed polysemy, also in different types of text (Louwerse et al., 2004).
Observed polysemy was the focus of Gale et al. (1992) and Krovetz (1998), who report the proportion of repeated polysemous words used in more than one sense in a discourse. Barba et al. (2021) propose a measure called "expressed polysemy" consisting in the ratio of the observed vs the potential number of senses of each lexeme. Pasini and Navigli (2018) compute a sense distribution for each word, and from it they derive an entropy score. Scores can then be averaged over each unique word in a text, producing a measure of "observed sense dispersion". In this study, our focus is on observed polysemy, but we also include measures of potential polysemy for comparison. We introduce a new measure of observed polysemy that takes into account the number of senses in which words are used in a text independent of their potential number of senses.
One possible reason why observed polysemy has not been addressed as much, particularly for spoken text, is the lack of appropriate corpora manually annotated with lexical semantic information. The datasets available in English –the language we work with–, such as SemCor (Miller et al., 1993), contain mostly written, scripted, monolog-like language.[1] The only dataset we are aware of is Ontonotes (Hovy et al., 2006). Part of this corpus, including texts extracted from multi-party talk shows, is annotated with word sense. We, however, do not use it in our study because only nouns and verbs are annotated, and we choose to focus on dyadic conversations. Given the lack of manually annotated datasets, we use ESCHER, a recent, state-of-the-art WSD system (Barba et al., 2021) to automatically annotate the corpora used in our study.

2.2. The OSD Hypothesis

Gale et al. (1992) observed that when a polysemous word is used multiple times in a text, it is "extremely" likely (94%) that all instances of this word are used in the same sense. They formulated the "one sense per discourse" (OSD) hypothesis and proposed to use it as a constraint for WSD models. Numerous variations of this hypothesis have been proposed, such as "one sense per collocation" (Yarowsky, 1993), "one domain per discourse" (Magnini et al., 2002) "one translation per discourse" (Carpuat, 2009), and others (Martinez and Agirre, 2000; Gella et al., 2014; Scarlini et al., 2019; Hauer and Kondrak, 2020). The heuristic has been used for WSD and related tasks such as Named Entity Recognition (Cucerzan and Yarowsky, 1999; Cucerzan, 2007; Wu and Giles, ; Pilehvar and Navigli, 2015; Chaplot and Salakhutdinov, 2018) or Machine Translation (Ture et al., 2012). It also has, however, been questioned.
In Gale et al. (1992)'s experiments, subjects were asked to judge, for a context pair, whether two word instances were used in the same sense. The agreement rate was very high, at 96.8%. Krovetz (1998) notes that the conclusions raised with this methodology are likely to involve mainly homonymy (i.e., unrelated word senses), since differences between homonyms are easier to agree upon. The author carries out a similar analysis based on finer-grained WordNet distinctions on two sense annotated corpora, SemCor and DSO (Ng and Lee, 1996) and finds that, overall, 33% of polysemous words are used in more than one sense within a discourse, as opposed to only 6%. Magnini et al. (2002) replicate Krovetz (1998)'s experiments with coarser-grained senses (WordNet domains, (Magnini and Cavaglià, 2000)) and find only 10% of the words to deviate from their "one domain per discourse" hypoth-

[1] We note that some SemCor texts contain parts of dialog (Agirre and Rigau, 1996), but these are not easily identifiable.

esis. Another criticism to OSD stems from the limited sample size: the original experiment relied on only 54 context pairs of nine polysemous nouns. Leacock et al. (1998) specifically point out that verbs and adjectives have a stronger tendency of being "nontopical", i.e., they have senses that are topic-independent and therefore may be found across texts on different topics, or used in different senses within the same document.
What these and other studies that address observed polysemy (McCarthy et al., 2007; Pasini and Camacho-Collados, 2020; Barba et al., 2021) have in common is that they are limited to written, monolog-like texts. Despite the relevance and pervasiveness of the spoken modality, a study of the polysemy level in spoken language is missing, and our work aims to fill this gap. We use WordNet synsets as well as coarser-grained senses, and consider a large set of words: all nouns, verbs and adjectives in a discourse.

3. Data

We select a number of datasets which allow us to compare texts along three different axes: spoken/written, dialog/monolog, scripted/spontaneous. In Table 1 we position datasets along each axis. We extract several of them from the SILICONE benchmark (Chapuis et al., 2020). When a dataset has a train/validation/test split, we include discourses from all subsets. We only consider discourses of over 100 words. We analyse the following spoken language corpora:

- **Debate**. We use the 2020 US presidential debate between Joe Biden and Donald Trump.[2] The debate is organized in topics which are proposed by the moderator. Each candidate has their say on the subject and then they engage in a dialog. We use the longer (typically topic-initial) interventions as monologs and the subsequent interactions as dialogs. The debate format is very convenient for our analysis, as it allows us to make a direct comparison between monologs and dialogs relating to the same topics.

- The **Iemocap** (Busso et al., 2008) database includes 151 dyadic interactions between actors. Conversations involved hypothetical situations eliciting specific emotions. We split the dataset into two sets, according to whether conversations were improvised or scripted.

- The **JUSThink** corpus (Norman et al., 2021) contains conversations between children aged 9 to 12 engaging in a collaborative learning activity on graphs which is presented by a robot. Each child needs to answer a test before and after this exercise. The results of this test are used to calculate the learning outcome, which can be interpreted as an indication of task success. We use the transcripts that are available for 10 conversations; and also the the calculated learning outcomes to investigate their relationship with polysemy.

- **Switchboard** (Stolcke et al., 2000) consists of 1,126 short dyadic conversations on a provided topic.

- The **HCRC MapTask Corpus** (Thompson et al., 1993)[3] contains 128 dyadic task-oriented conversations. Participants needed to collaborate in order to re-draw, on a map, a route that was only visible to one interlocutor.

- **BT Oasis corpus** (Leech and Weisser, 2003)[4] is a collection of 635 (378 after filtering for text length) calls made to the British Telecom and Trainline operator services.

	Spoken			Written
	Scripted	**Spontaneous**		**Scripted**
Dialog	Iemocap (scripted)	Iemocap (impro)	JUSThink	
		Switchboard	Oasis	
		Debate (dialog)	MapTask	
Monolog		Debate (mono.)		Senseval-2
				Senseval-3
				SemEval 2015

Table 1: Dataset classification along the three axes.

As instances of written text we analyze data from three WSD evaluation campaigns. The reason for including these is that they contain manual sense annotations. This allows us to check the validity of the polysemy estimations made based on automatic sense annotations. We obtain these datasets from the unified WSD Framework (Raganato et al., 2017), where all sense annotations have been mapped to the WordNet 3.0 inventory.

- **Senseval-2** (Edmonds and Cotton, 2001) consists of three texts on traditions, medical research, and education.

- **Senseval-3 task 1** (Snyder and Palmer, 2004) also contains three documents: an editorial and a news story from WSJ and a fictional story from the Brown corpus.

- **SemEval-15 task 13** (Moro and Navigli, 2015). The data for this task comprises four documents on three different domains: biomedical (two texts), mathematics/computing and social issues.

We exclude from our study the data from other campaigns, SemEval-07 task 17 (Pradhan et al., 2007) and SemEval-13 task 12 (Navigli et al., 2013), because they

[2] https://www.kaggle.com/rmphilly18/us-presidential-debatefinal-october-2020

[3] https://groups.inf.ed.ac.uk/maptask/transcripts/

[4] Obtained from https://github.com/NathanDuran/BT-Oasis-Corpus.

do not contain annotations for all parts of speech included in our analyses. We did not use SemCor because the WSD model that we use for annotation was trained on this corpus.

4. Methodology

In this Section we present our approach. We first perform word sense disambiguation on the corpora presented above, and we calculate the polysemy level of each dataset using different measures of polysemy and definitions of "sense".

4.1. Automatic Word Sense Annotation

We automatically annotate every dataset described in Section 3 with senses using a state-of-the-art WSD system, ESCHER (Barba et al., 2021). This model, trained on SemCor, relies on a Transformer-based architecture and a reformulation of the typical WSD objective. The model needs to select the gloss corresponding to the meaning of a target word in a sentence. We first apply tokenization, lemmatization and part-of-speech (PoS) tagging to every text with the `stanza` package (Qi et al., 2020).[5] We feed isolated sentences into the WSD model and obtain the word senses of content words in it. In our polysemy analysis, we only include nouns, verbs and adjectives.

4.2. Polysemy Measures

We present below the measures of polysemy used in our analysis.[6] We use two measures of potential polysemy (PA and PCT-POLY), two of observed polysemy (MOSD and AVGSENSES) and a measure of observed sense dispersion (ENTROPY). We use NP_w to refer to the number of *potential* senses a word w (a lemma with a specific part of speech)[7] can have according to WordNet, and $NO_{w,t}$ for the number of senses w has been *observed* with in text t.

Potential ambiguity (PA) PA consists in the average number of senses of word instances in text t according to WordNet:

$$PA(t) = \frac{\sum_{w \in V_t} (NP_w \times c(w,t))}{\sum_{w \in V_t} c(w,t)} \quad (1)$$

where V_t is the vocabulary of text t, and $c(w,t)$ is the frequency –the number of occurrences– of w in t.

[5]We skip this step for the Senseval and SemEval corpora, as it is already done by Raganato et al. (2017).

[6]The code for calculating these measures will be available at: `https://github.com/ainagari/spoken_poly`.

[7]We use *word* to refer to a vocabulary entry consisting of a lemma paired with a part-of-speech. We refer to one occurrence (or token) as a (*word*) *instance*.

Percentage of polysemous words (PCT-POLY) We calculate the proportion of polysemous word instances out of all word instances in a text t. In Equation 2, V_t^{NP+} is the subset of words in V_t that are polysemous (i.e., that have more than one sense in WordNet, $NP_w > 1$).

$$\text{PCT-POLY}(t) = \frac{\left(\sum_{w \in V_t^{NP+}} c(w,t) \right) \times 100}{\sum_{w \in V_t} c(w,t)} \quad (2)$$

More than One Sense per Discourse (MOSD) As in Krovetz (1998), we calculate the percentage of repeated polysemous words used in more than one sense in a text. Formally,

$$\text{MOSD}(t) = \frac{\left| \left\{ w \in V_t^{NP+} : c(w,t) > 1 \right. \right. \\ \left. \left. \wedge NO_{w,t} > 1 \right\} \right| \times 100}{|\{w \in V_t^{NP+} : c(w,t) > 1\}|} \quad (3)$$

AVGSENSES The MOSD metric takes into account the number of words observed in more than one sense in t, but it ignores the number of different senses in which words are used. We propose AVGSENSES, which calculates the average number of senses in which words in a discourse are used. It is similar to PA, but the number of senses of a word is the observed one and not the potential one taken from WordNet. We determine, for every word, the number of different senses s it has been used with in t. We then count all instances of words having exactly s observed senses. We calculate AVGSENSES as an average of these grouped data:

$$\text{AVGSENSES}(t) = \frac{\sum_{s=1}^{max(NO_{*,t})} \left(s \times \sum_{w \in V_t^{NO=s}} c(w,t) \right)}{\sum_{w \in V_t} c(w,t)} \quad (4)$$

where $max(NO_{*,t})$ is the highest number of senses found for any word in text t. $V_t^{NO=s}$ is the set of words in text t that have been used in s different senses.

ENTROPY The measures presented so far are not sensitive to how senses are distributed. However, this is something useful to consider: it gives us a more complete view of the ambiguity brought by a polysemous word, which will be felt differently if there is a largely predominant sense or multiple senses that are in the same order of importance.

Following Pasini and Navigli (2018), we calculate an entropy-based measure of word sense distribution. The authors use a WSD model to compute, for every word

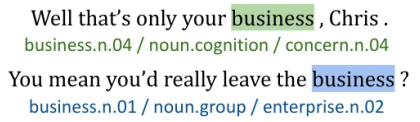

Figure 1: Example of automatic annotation on Iemocap with different sense granularities.

instance, a probability distribution over the word's possible senses. In our case, these distributions have all their probability allocated in the single best sense proposed by ESCHER for that word instance. For every word w in V_t used more than once in text t ($w \in V_t^{c>1}$), we compute its sense distribution in t, $sd_{w,t}$, by adding the probabilities of each sense through all instances of w in t and normalizing them by the sum of all probabilities. The ENTROPY of t is then calculated as the average entropy[8] $\mathcal{H}$ of words in t:

$$\text{ENTROPY}(t) = \frac{\sum_{w \in V_t^{c>1}} \mathcal{H}(sd_{w,t})}{|V_t^{c>1}|} \quad (5)$$

4.3. Granularity of the Sense Inventory

As noted by Krovetz (1998) and Magnini and Cavaglià (2000), the conclusion of the "(more than) one sense per discourse" hypothesis depends on the granularity of the sense inventory used. Naturally, a fine-grained inventory like WordNet will generally result in estimations of higher polysemy than coarser-grained senses. In order to verify whether the hypothesis holds for different kinds of texts and different granularities, we experiment with three definitions of sense, or *sense types*:

- WordNet **synsets**. These are automatically annotated by the WSD model;
- WordNet **supersenses**. Supersenses are 26 categories which classify all noun, verb and adjective synsets in WordNet.[9]
- WordNet **hypernyms**. We use WordNet's direct hypernym of a synset. For example, two senses of *table* (noun) share a hypernym synset, *furniture*.[10] Adjectives in WordNet do not have hypernyms. Consequently, they are treated like words that are missing in WordNet and are omitted from this analysis.

When using a sense type other than synsets, we determine whether a word is polysemous based on that sense type. As a consequence, the number of ambiguous words in a text V_t^{NP+} varies depending on the sense type used. Figure 1 shows two word instances annotated with all sense types. We report results using synsets, as done in previous work (Krovetz, 1998), and explain how we choose the best alternative sense type in Section 5.1.2.[11]

4.4. Dataset Comparison

We run three main comparisons. We use the Debate and the Iemocap datasets to investigate the dialog-monolog and scripted-spontaneous distinctions, respectively. We also compare the polysemy levels of texts in each of the four categories represented in Table 1. To calculate the polysemy level of a dataset, we simply average the polysemy values obtained for every text. For a text category (e.g., spontaneous spoken dialog), we average the polysemy values obtained for each dataset in it.

5. Methodology Validation

The proposed methodology relies on the assumption that the automatically annotated senses are sufficiently reliable. Here, we discuss this assumption (Section 5.1). We also consider the relation between the polysemy measures and other text-level properties (Section 5.2).

5.1. Automatic Sense Annotation

Manual annotation of word senses involves an expensive and time-consuming effort. As a consequence, the number of corpora with this kind of information is limited. Researchers have resorted to automatic sense annotation and similar strategies to create large sense-annotated corpora (Delli Bovi et al., 2017; Scarlini et al., 2019), which have been useful as training data for WSD models (Pasini and Navigli, 2017). Automatically annotated data, however, are inevitably noisy and may contain errors. We account for this limitation in two ways: through the manual validation of the WSD model's output, and by comparing the polysemy rankings obtained with automatic and manual annotations.

5.1.1. Annotation Quality

To assess the quality of the automatic annotation, we evaluate the predictions made by ESCHER on the Senseval and SemEval corpora, for which manual annotations are available. For every instance of a polysemous noun, verb and adjective, we check whether the annotation proposed by the model corresponds to the gold sense. When an instance is manually annotated with multiple senses, we count the predicted sense as correct if it is among them. The percentages of correctly

[8] This *average* entropy, being a combination of scores over different distributions, is not interpretable in the usual information-theoretic sense. It, however, provides a useful insight on a discourse's observed ambiguity.

[9] Most adjective synsets are assigned the same supersense: only 6% of all adjectives in WordNet are considered polysemous under this sense type.

[10] When a synset has multiple hypernyms (which is rare), we treat a set of hypernyms as a sense, distinct from the hypernyms it is made of.

[11] Using hypernyms reduces the potential polysemy of 23% of all unique polysemous noun and verb lemmas present in our datasets. With supersenses, polysemy is reduced for 71% of all unique lemmas.

	Sense type	MOSD	AVGSENSES	ENTROPY
Discourses	synsets	**0.644**	**0.644**	**0.644**
	supersenses	0.455	0.511	**0.644**
	hypernyms	0.422	0.556	0.466
Datasets	synsets	1	1	1
	supersenses	1	1	0.333
	hypernyms	1	1	1

Table 2: Kendall's τ correlation between rankings obtained with manual vs automatic annotations in the Senseval and SemEval datasets. We compare the rankings of the ten discourses (top) and the three datasets (bottom).

annotated instances are fairly high: 77.1%, 73.4% and 79.9% in Senseval-2, Senseval-3 and SemEval-2015. The number of instances is 1539, 1673 and 800, respectively.

The WSD model used for automatic disambiguation has been trained and evaluated on scripted written monologs. Barba et al. (2021) note that the sentences where ESCHER makes mistakes are, on average, shorter than the average sentence in their test data. Notably, sentences in the dialog corpora used in our study are overall shorter than those in the monolog data.[12] It is, therefore, likely that its accuracy on dialog data is lower than on written or monolog data. We carry out a manual validation of the model's predictions on texts of each category to estimate its performance. We pick five texts of manageable length: three spontaneous spoken dialogs from different datasets, one scripted dialog and one spoken monolog. Complete results and details of this validation are included in Appendix A. We observe that the accuracy scores, ranging from 75.0% (Maptask) to 88.5% (Iemocap-impro) are comparable to results on written monolog. This result is encouraging, as it shows that the model can perform well on dialog and spoken data, but the sample is too small to consider it a reliable estimation of its accuracy on the corresponding discourse types.

5.1.2. Validation of Polysemy Estimations

Due to noise in the automatic annotations, the observed polysemy estimations we obtain will not be exact and need to be interpreted with caution. Our goal, rather than calculating a precise polysemy value for a text, is to compare the relative polysemy estimations of texts and determine which are more polysemous than others. We want our estimations to provide a ranking of texts by polysemy which is faithful to the ranking we would obtain with manual annotations. We verify that on Senseval-2, Senseval-3 and Semeval-2015. Specifically, we calculate the polysemy level of all discourses in these datasets with the different measures of observed polysemy[13] and sense definitions (see Sections 4.2 and 4.3). We then compare the rankings obtained by each measure with the manual vs. the automatic annotation by means of Kendall's τ. A higher τ indicates a higher similarity, and a higher reliability of the rankings obtained with automatic annotations. We do the same at the dataset level. Results are shown in Table 2. At the discourse level, agreement between rankings is moderate, ranging from 0.422 to 0.644. With regard to the different sense types, synsets obtain the highest correlation with the three measures ($\tau = 0.644$). Supersenses are overall the second best sense type. When aggregating the polysemy levels of the three datasets, all settings but one[14] produce the same ranking for the two types of annotation ($\tau = 1$). Overall, the rankings obtained with the two kinds of annotation are fairly similar. In what follows, we will report results using synsets as well as supersenses as the most reliable sense types.

Measure	Text Length	LexDiv
PA	-0.04	-0.13
PCT-POLY	0.04	-0.06
MOSD	0.45	-0.32
AVGSENSES	0.79	-0.73
ENTROPY	0.49	-0.32

Table 3: Spearman's ρ correlation of each polysemy measure with text length and lexical diversity.

5.2. Discourse Length and Lexical Diversity

The observed polysemy measures used in our analyses are likely to be affected by discourse length: in longer texts words may be reused more often, which increases their chances of being used in multiple senses. The first column in Table 3 shows the correlations found between discourse length (calculated as the number of tokens in a text) and each polysemy measure (with synsets).We indeed find a strong correlation with AVGSENSES ($\rho = 0.79$) and moderate correlations with MOSD and ENTROPY. The longer the text, the less likely OSD is to hold. This is important for WSD, and also for our results, which should be interpreted in the light of this text-level property. The measures of potential polysemy do not correlate with text length.

We also explore the relationship between lexical diversity (LexDiv) and polysemy. We define the LexDiv of a text t as the number of unique words ($|V_t|$) divided by the number of tokens in a text ($|t|$). For this metric, we do not restrict the POS of lemmas in V_t. We hypothesize that texts with a richer lexical diversity have a lower observed polysemy level, because they contain fewer repetitions. We, in turn, expect this measure of lexical diversity to be negatively correlated with discourse length (McCarthy and Jarvis, 2010). Results are

[12] Average sentence length for spnt. spoken dialog: 8.8; scripted spoken dialog: 8.8; spnt. spoken monolog: 17.26; scripted written monolog: 21.08.

[13] Measures of potential polysemy are not included in this comparison because they do not require sense annotations.

[14] The low correlation of ENTROPY with supersenses is due to a small numerical difference (0.01) between two datasets.

Dataset	# texts	Avg Text Length	Avg LexDiv
Debate (dialog)	6	2165.3	0.22
Iemocap (impro)	80	919.1	0.24
JUSThink	10	2950.6	0.10
Switchboard	1126	1704.1	0.20
MapTask	128	1287.3	0.20
Oasis	378	340.2	0.33
Iemocap (scripted)	71	1073.7	0.28
Debate (mono.)	15	316.1	0.41
Senseval-2	3	1922.0	0.33
Senseval-3	3	1847.0	0.33
SemEval-2015	4	651.0	0.37

Table 4: Average discourse length and lexical diversity per dataset.

shown in the second column of Table 3. We again find a strong correlation with AVGSENSES, this time negative ($\rho = -0.73$); and moderate negative correlations with the other measures of observed polysemy. LexDiv and discourse length are also, as expected, strongly correlated ($\rho = -0.82$).

Table 4 shows the average text length and LexDiv values for each dataset. We observe that spoken and written monologs (the Debate, Senseval and SemEval datasets) have an overall larger lexical diversity. This indicates that, in dialogs, words are generally reused more often, which is associated with higher levels of observed polysemy. Interestingly, we also see that Iemocap (impro) has a lower lexical diversity than Iemocap (scripted).

6. Results

In this Section, we present the results of our analysis. We compare the polysemy of different discourse types (Section 6.1) and investigate the correlation between polysemy and learning outcome (Section 6.2).

6.1. Polysemy Level of Different Text Types

Table 5 shows the polysemy values obtained for each dataset and discourse type (colored rows), using synsets (on the left part) and supersenses (on the right).

Monolog vs dialog When comparing the results obtained for Debate-monologs and Debate-dialogs, we see that, according to MOSD and AVGSENSES, dialogs have a higher observed polysemy than monologs. The reverse is true when considering the measures of potential polysemy (PA and PCT-POLY), but differences between these are small. ENTROPY values are also quite similar between the two datasets. This indicates that monologs contain a slightly higher proportion of polysemous words, which have on average a higher number of senses. Polysemous words are, however, used in more senses in dialogs. This observation with respect to the observed polysemy may partly be due to the difference in average text length between the two datasets (see Table 4). The datasets, however, are comparable in terms of topics and speakers. Spoken monologs are known to be harder than dialog in terms of production (Pickering and Garrod, 2004), so it is natural that they are shorter than comparable dialogs.

Scripted vs spontaneous We compare the rankings obtained for the scripted and improvised parts of Iemocap. In this case, the two datasets have a comparable text length. Spontaneous texts have higher PA, PCT-POLY and AVGSENSES than scripted texts, but a lower MOSD and ENTROPY. This contrast between the observed polysemy measures (MOSD and AVGSENSES) suggests a tendency for spontaneous texts to contain a smaller proportion of words used in multiple senses than scripted texts, but those words are reused in a higher number of different senses in improvised conversation. The PA values show that words used in spontaneous texts have, overall, more senses than words in scripted texts. We believe this has to do with word frequency: higher frequency is linked to a higher number of (potential) senses (Zipf, 1945). We find that in spontaneous texts, words are indeed more frequent than in scripted ones.[15]

All discourse types The polysemy values obtained for each of the four text types are found in the colored rows. This comparison reveals that scripted written monologs are the least polysemous kind of discourse according to all our definitions of text polysemy and to the two sense types. Note that this result cannot be attributed to text length (see Table 4). Omitting the dataset with shortest texts in this category, SemEval-2015, would not affect the placement of this type of text in the rankings. Interestingly, doing so would bring MOSD up to 32.2%, very close to Krovetz (1998)'s estimations on SemCor (33%).

The other text types, all consisting of spoken language, have a higher polysemy than written discourse. Importantly, they present, for most measures, clearly distinct polysemy levels. This confirms our general hypothesis that texts of different nature have different polysemy levels. Not all polysemy measures rank the text types in the same order, but we do see that dialogs tend to have a higher observed polysemy (MOSD and AVGSENSES) than spoken monologs. This is, however, not true of the Oasis dataset, which is comparable to Debate-monologs in terms of average length. We cannot, therefore, be certain that this difference is due to the type of text.

It is worth noting the variability between datasets even within a specific type, particularly in spontaneous spoken dialog. For instance, in this category, MOSD (synsets) ranges between 23.1 (Oasis) and 47.2 (JUSThink). This is because there are other factors that may play a role on the polysemy level of a text, such as its

[15] We calculate the frequencies of noun, verb and adjective instances in each discourse with Python's `word_freq` package. Average frequencies are 217.3 and 343.3 per million occurrences in scripted and improvised texts.

	Synsets					Supersenses				
Dataset	PA	PCT-POLY	MOSD	AVGSENSES	ENTROPY	PA	PCT-POLY	MOSD	AVGSENSES	ENTROPY
Debate (dialog)	10.3	93.1	37.8	1.76	0.26	3.9	74.0	**41.8**	1.43	0.14
Iemocap (impro)	**11.3**	**93.7**	39.9	1.63	0.28	4.2	76.1	29.2	1.34	0.16
JUSThink	10.8	**93.7**	**47.2**	**2.53**	0.31	4.4	**84.2**	40.0	**2.02**	**0.20**
Switchboard	9.8	91.5	39.2	1.68	0.27	3.8	72.5	29.5	1.37	0.15
MapTask	11.3	89.3	35.5	1.62	0.21	**4.7**	75.5	30.0	1.43	0.14
Oasis	10.3	89.9	23.1	1.27	0.14	4.0	73.3	16.9	1.16	0.08
Avg (spt. spk. dialog)	10.6	91.9	37.1	1.75	0.24	4.2	75.9	31.2	1.46	0.14
Iemocap (scripted)	9.9	93.1	44.1	1.51	**0.32**	3.7	71.1	29.7	1.27	0.16
Debate (mono.)	10.9	93.6	33.1	1.28	0.23	4.0	72.3	36.6	1.19	0.14
Senseval-2	6.4	84.5	30.1	1.20	0.17	2.9	65.3	21.1	1.10	0.09
Senseval-3	7.6	84.6	34.3	1.26	0.21	3.3	67.7	27.3	1.16	0.12
SemEval-2015	6.6	86.1	20.1	1.12	0.12	3.1	72.5	20.5	1.09	0.10
Avg (scr. wrt. mono.)	6.9	85.1	28.2	1.19	0.17	3.1	68.5	22.9	1.12	0.10

Table 5: Polysemy levels of each dataset using different metrics and sense types. The highest polysemy value obtained by each measure is in bold, and the lowest is underlined. Colored rows correspond to the values obtained for a discourse type.

topic, formality or lexical complexity. For example, texts on specialized topics may contain a higher proportion of rare words, which tend to be monosemous (Zipf, 1945). The comparisons run within a corpus (on the Debate and Iemocap datasets) allow for minimizing the effect of these variables, as topics and style are similar throughout the corpus. When aggregating multiple datasets under one category, we have less control over these differences. Still, this variation should discourage generalizations like OSD and instead motivate a more individualized approach to disambiguation where the properties of each kind of text are taken into account.

When using supersenses, we still observe a prominent proportion of ambiguous words used in multiple senses, ranging from 20.5 (SemEval-2015) to 41.8 (Debate-dialogs). Thus, even when considering coarser-grained senses, the original OSD estimations of 6% (Gale et al., 1992) are probably too strict.

We calculate the correlations between polysemy measures. The strongest correlations found are between ENTROPY and the other observed polysemy measures (0.92 with MOSD and 0.65 with AVGSENSES). Correlations between observed and potential polysemy are weak ($0.12 \leq \rho \leq 0.24$). This highlights the difference between the two notions of polysemy.

6.2. Relationship with Learning Outcome

One interesting question we can explore with our analysis is whether the polysemy level of a dialog affects, or is related to, its successful development. We hypothesize that the polysemy level of a dialog may reflect speakers' mutual understanding. For example, a lower polysemy level could indicate that dialog participants are on the same page or agree with each other.

As a first step, in this study we calculate the Spearman's correlation between the polysemy level (synsets) of the ten conversations in the JUSThink dataset with the learning outcome (LO) of its participants (see Section 3). We find weak correlations with PA ($\rho = 0.26$), MOSD ($\rho = -0.24$) and AVGSENSES ($\rho = -0.32$). Correlations are neither strong nor significant, due in part to the small sample size. We can see that dialogs with a higher LO have a slight tendency to exhibit lower observed polysemy, which is in accordance with our expectations. The correlation with PA, however, is positive: using words with a higher number of senses tends to result in a higher learning outcome. We also find a weak positive correlation between LO and text length ($\rho = 0.29$). In longer texts, participants have the possibility to learn more. Nevertheless, correlations are not strong enough to extract definite conclusions. PCT-POLY and ENTROPY do not correlate with LO ($|\rho| < 0.10$).

7. Conclusion and Future Work

We investigate and compare, for the first time, the polysemy level of different kinds of texts; notably, of spoken and conversational language. We show that the use of polysemous words changes depending on the nature of the discourse. Specifically, spoken texts tend to have a higher polysemy level than written discourses. This holds for different views of polysemy; potential and observed.

We believe that this study opens exciting avenues for future work. The relationship between polysemy and dialog success is worth investigating further, with more data and different definitions of success (for example, based on the appreciation of mutual understanding). Another interesting direction to pursue would be identifying and characterizing words that tend to be used in more senses throughout a discourse. With regard to dialog, we intend to investigate speaker alignment and possible misunderstandings by modeling the introduction of new senses into a conversation by each speaker. Finally, an obvious continuation of this work would involve including other text types, such as spontaneous written dialogs; or broadening the context that the WSD model uses for disambiguation.

We hope that our study will motivate further work on text-level polysemy, and that the insights provided here will be useful for the disambiguation of different kinds of texts.

8. Acknowledgments

We thank the anonymous reviewers for their helpful feedback and suggestions. This work was supported by the Télécom Paris research chair on Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS).

9. Bibliographical References

Agirre, E. and Rigau, G. (1996). Word Sense Disambiguation using Conceptual Density. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Barba, E., Pasini, T., and Navigli, R. (2021). ESC: Redesigning WSD with Extractive Sense Comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online, June. Association for Computational Linguistics.

Bazzanella, C. and Damiano, R. (1999). The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, 31(6):817–836.

Busso, C. and Narayanan, S. S. (2008). Scripted dialogs versus improvisation: lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Proc. Interspeech 2008*, pages 1670–1673.

Carpuat, M. (2009). One Translation Per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, Colorado, June. Association for Computational Linguistics.

Chaplot, D. S. and Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. In *Proc. of the 32th AAAI Conference on Artificial Intelligence*, New Orleans, USA.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication.

Cucerzan, S. and Yarowsky, D. (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

Delli Bovi, C., Camacho-Collados, J., Raganato, A., and Navigli, R. (2017). EuroSense: Automatic Harvesting of Multilingual Sense Annotations from Parallel Text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 594–600, Vancouver, Canada, July. Association for Computational Linguistics.

Friedberg, H., Litman, D., and Paletz, S. B. (2012). Lexical entrainment and success in student engineering groups. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 404–409. IEEE.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One Sense Per Discourse. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Garrod, S. and Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Gella, S., Cook, P., and Baldwin, T. (2014). One Sense per Tweeter ... and Other Lexical Semantic Tales of Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 215–220, Gothenburg, Sweden, April. Association for Computational Linguistics.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Hauer, B. and Kondrak, G. (2020). One Homonym per Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

Kaur, J. (2011). Intercultural communication in english as a lingua franca: Some sources of misunderstanding. *Intercultural Pragmatics*, 8(1):93–116.

Krovetz, R. (1998). More than One Sense Per Discourse. *NEC Princeton NJ Labs., Research Memorandum*, 23.

Leacock, C., Chodorow, M., and Miller, G. A. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. *Computational Linguistics*, 24(1):147–165.

Liberman, K. (2012). Semantic drift in conversations. *Human Studies*, 35(2):263–277.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., and Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In *K.Forbus, D. Gentner & T.Regier (Eds.), Proceedings of the twenty-sixth annual conference of the Cognitive Science Society*, pages 843–848. Mahwah, NJ: Erlbaum.

Magnini, B. and Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Magnini, B., Strapparava, C., Pezzulo, G., and Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4):359–373.

Martinez, D. and Agirre, E. (2000). One Sense per

Collocation and Genre/Topic Variations. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 207–215, Hong Kong, China, October. Association for Computational Linguistics.

McCarthy, D. and Carroll, J. (2003). Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences. *Computational Linguistics*, 29(4):639–654, 12.

McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2007). Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.

Nerlich, B. and Clarke, D. D. (2001). Ambiguities we live by: Towards a pragmatics of polysemy. *Journal of Pragmatics*, 33(1):1–20.

Pasini, T. and Camacho-Collados, J. (2020). A Short Survey on Sense-Annotated Corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5759–5765, Marseille, France, May. European Language Resources Association.

Pasini, T. and Navigli, R. (2017). Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, Copenhagen, Denmark, September. Association for Computational Linguistics.

Pasini, T. and Navigli, R. (2018). Two Knowledge-based Methods for High-Performance Sense Distribution Learning. In *Proc. of the 32th AAAI Conference on Artificial Intelligence*, New Orleans, USA.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Pickering, M. J. and Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, 4(2):203–228.

Pilehvar, M. T. and Navigli, R. (2015). From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artificial Intelligence*, 228:95–128.

Preiss, J. and Stevenson, M. (2013). Unsupervised Domain Tuning to Improve Word Sense Disambiguation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 680–684, Atlanta, Georgia, June. Association for Computational Linguistics.

Pronin, E., Puccio, C., and Ross, L. (2002). Understanding misunderstanding: social psychological perspectives. In *In T. Gilovich, D. Griffin, & D. Kahneman, Eds., Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July. Association for Computational Linguistics.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

Scarlini, B., Pasini, T., and Navigli, R. (2019). Just "OneSeC" for Producing Multilingual Sense-Annotated Data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709, Florence, Italy, July. Association for Computational Linguistics.

Sinha, T. and Cassell, J. (2015). We Click, We Align, We Learn: Impact of Influence and Convergence Processes on Student Learning and Rapport Building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence*, INTERPERSONAL '15, page 13–20, New York, NY, USA. Association for Computing Machinery.

Ture, F., Oard, D. W., and Resnik, P. (2012). Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 417–426, Montréal, Canada, June. Association for Computational Linguistics.

Wu, Z. and Giles, C. L.). Sense-aware semantic analysis: A multi-prototype word representation model using wikipedia. In *Twenty-ninth AAAI conference on artificial intelligence*, Austin, TX, USA.

Yarowsky, D. (1993). One Sense per Collocation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, 33(2):251–256.

10. Language Resource References

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Chapuis, E., Colombo, P., Manica, M., Labeau, M., and Clavel, C. (2020). Hierarchical Pre-training for

Sequence Labelling in Spoken Dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online, November. Association for Computational Linguistics.

Edmonds, P. and Cotton, S. (2001). SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, July. Association for Computational Linguistics.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (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.

Leech, G. and Weisser, M. (2003). Generic speech act annotation for task-oriented dialogues. In *Proceedings of the corpus linguistics 2003 conference*, volume 16, pages 441–446. Citeseer.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Moro, A. and Navigli, R. (2015). SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

Navigli, R., Jurgens, D., and Vannella, D. (2013). 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. Association for Computational Linguistics.

Ng, H. T. and Lee, H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47, Santa Cruz, California, USA, June. Association for Computational Linguistics.

Norman, U., Dinkar, T., Bruno, B., and Clavel, C. (2021). Studying Alignment in Spontaneous Speech via Automatic Methods: How Do Children Use Task-specific Referents to Succeed in a Collaborative Learning Activity? *arXiv preprint arXiv:2104.04429*.

Pradhan, S., Loper, E., Dligach, D., and Palmer, M. (2007). SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June. Association for Computational Linguistics.

Snyder, B. and Palmer, M. (2004). The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Thompson, H. S., Anderson, A., Bard, E. G., Doherty-Sneddon, G., Newlands, A., and Sotillo, C. (1993). The HCRC Map Task Corpus: Natural Dialogue for Speech Recognition. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

A. Manual Validation of Annotation Quality: Results

As explained in Section 5.1.1, we manually verify the quality of the automatic annotation on dialogs and spoken monologs on a sample of texts from different datasets. Table 6 contains details about these texts and the annotation accuracy on each of them. In the spontaneous Iemocap dialog, 35 instances correspond to the expressions "you know" and "I mean", which were all considered to be annotated correctly. Excluding them, the accuracy would be 84.4%.

Category	Dataset	Tokens	Annotations	Acc
spt. dialog	Iemocap	589	131	88.5
	Switchboard	772	160	77.5
	Maptask	867	224	75.0
scr. dialog	Iemocap	530	118	76.3
spt. mono.	Debate	451	118	75.4

Table 6: Details and annotation accuracy of the selected spoken texts. The number of annotations includes only polysemous nouns, verbs and adjectives. *Acc* is the percentage of sense annotations that were judged to be correct. The IDs of the texts in their original datasets are: Ses04F_impro06 and Ses04M_script01_2 (Iemocap), 13 (Switchboard), q2ec2 (Maptask), 3 (Debate).