# Towards Human Evaluation of Mutual Understanding in Human–Computer Spontaneous Conversation: An Empirical Study of Word Sense Disambiguation for Naturalistic Social Dialogs in American English

**Alex Lưu**
Brandeis University
alexluu@brandeis.edu

## Abstract

Current evaluation practices for social dialog systems, dedicated to human–computer spontaneous conversation, exclusively focus on the quality of system-generated surface text, but not human-verifiable aspects of mutual understanding between the systems and their interlocutors. This work proposes Word Sense Disambiguation (WSD) as an essential component of a valid and reliable human evaluation framework, whose long-term goal is to radically improve the usability of dialog systems in real-life human–computer collaboration. The practicality of this proposal is proved via experimentally investigating (1) the WordNet 3.0 sense inventory coverage of lexical meanings in spontaneous conversation between humans in American English, assumed as an upper bound of lexical diversity of human–computer communication, and (2) the effectiveness of state-of-the-art WSD models and pretrained transformer-based contextual embeddings on this type of data.[1]

## 1 Introduction

As surveyed in Finch and Choi (2020), current evaluation practices for human–computer spontaneous conversation, including open domain dialog systems and chatbots, exclusively focus on the quality of system responses, e.g. how well the responses match ground truth human responses (based on certain automated metrics) or whether they are on-topic with the immediate dialog history (judged by a human). These evaluation practices potentially drive researchers into the race of generating better surface text while undermining or ignoring the ultimate goal of capturing mutual understanding between the systems and humans throughout the conversation (cf. the Great Misalignment Problem raised by Hämäläinen and Alnajjar, 2021). Consequently, current systems are unable to effectively function in real-life human–computer collaboration tasks. For example, the lack of genuine conceptual alignment with users leads to language learning chatbots being used only as reactive systems, even though theoretically they could provide the learners with the opportunity for free and flexible meaningful conversation (Bibauw et al., 2019), and consequently play a key role in supporting autonomous language learning beyond the classroom. To improve the usability of dialog systems for human–computer spontaneous conversation, their evaluation should include human-verifiable aspects of language competence which facilitate mutual understanding (instead of treating them as black box functions). Moreover, breaking down the evaluation into such concrete components would allow users' participation in system evaluation from early development stages (Heuer and Buschek, 2021).

Currently, talking to social chatbots without knowing which sense of a semantically ambiguous word[2] the chatbots have in their internal interpretation, human evaluators cannot identify the root cause of a problematic conversational move performed by the chatbots to provide more useful feeback. For example, examining the dialog shown in Figure 1, we can agree that the last utterance produced by the chatbot is not appropriate. However, we cannot know for sure if that is due to the chatbot's inadequate interpretation of "bank"[3] in the preceding question "What do you do at a river bank?", or its complete ignorance of the meaning of this word by just generating the most probable utterance according to the dataset it is trained on.

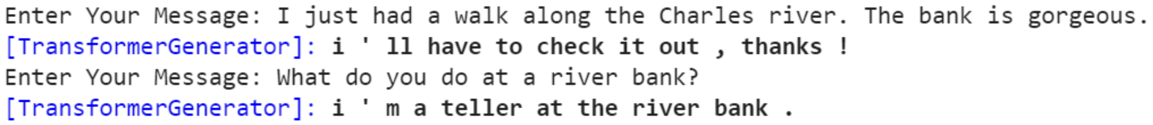
Arguably, one of the most natural ways for social chatbots to enhance the quality of their interaction with humans is explicitly assigning semantically ambiguous words specific senses, aka Word Sense Disambiguation (WSD), and using these senses

[1] The live version of this publication is located at https://osf.io/8u3gf/.

[2] Either polysemous or homonymous.

[3] As a financial institution instead of the land alongside a river, which is more felicitous in this particular context.

*Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 116 - 125
May 27, 2022 ©2022 Association for Computational Linguistics



```
Enter Your Message: I just had a walk along the Charles river. The bank is gorgeous.
[TransformerGenerator]: i ' ll have to check it out , thanks !
Enter Your Message: What do you do at a river bank?
[TransformerGenerator]: i ' m a teller at the river bank .
```

Figure 1: A dialog between me and a state-of-the-art (SOTA) chatbot developed by Meta Research (Roller).

for further reasoning[4] to demonstrate the chatbots' understanding capability with human-readable aspects of grounding (Clark, 1996) in the course of spontaneous conversation. This would improve human–computer communication in collaborative tasks by allowing the human partners to directly access the interpretable form of computers' model of conversation anytime they need to so that they can make adequate on-the-fly conversational adjustments. In addition, being able to access the computer's human-readable representation of conversational context in the evaluation regime, a human evaluator does not need to construct different interpretation alternatives and therefore can be confident that they are on the same page with other evaluators (cf. Appendix A – a small experiment that shows a wide divergence in human interpretation of a word token in spontaneous conversation). This transparency definitely reduces the subjectivity of the evaluation task, and therefore improves its reliability and reproducibility (Specia, 2021).

This work proposes and evaluates WSD as an essential component of a novel human evaluation framework intended for human–computer mutual understanding in spontaneous conversation in English, but also sensible for any tasks involving natural language interpretation. Specifically, based on the state of the art in WSD (Bevilacqua et al., 2021), it addresses the following research questions:

1. Can WordNet 3.0 (Fellbaum, 2010), the most popular English sense inventory, approximate word meaning in spontaneous dialog[5] well?
2. Are state-of-the-art (SOTA) WSD models, using transfer learning with both pretrained transformers and non-conversational sense-annotated data, ready for conversational text?
3. How effective is it to directly use contextual embeddings of pretrained transformers, e.g. BERT (Devlin et al., 2019) or its variants, to address WSD in spontaneous conversation?

The rationale behind (3) is to test the hypothesis that contextual embeddings of word tokens in spontaneous conversation are well correlated with definitions of their context-sensitive senses (versus task-oriented scenarios where the word senses are constrained by the task). When deploying a dialog system, the transparent integration of these embeddings with other components in the NLP pipeline is preferable over the "black box" nature of off-the-shelf end-to-end WSD models, which poses the challenges of how to (a) align these models' output with the system's NLP pipeline's, and (b) improve their real-time performance using knowledge about a specific instance of conversation.

To address (1–3), I first automatically annotated WordNet senses of ambiguous words in NEWT-SBCSAE, a publicly accessible corpus of naturally occurring spontaneous dialogs in American English (Lưu and Malamud, 2020; Riou, 2015; Du Bois et al., 2000), using both a SOTA WSD model and a simple baseline model directly based on contextual embeddings of pretrained transformers (Section 2.2). Next, I collected human judgements on the outputs of these models as well as the appropriate senses of the target words (Section 2.3). These judgments were then used to assess the coverage of the WordNet sense inventory (Section 3) and the efficacy of WSD models, including both models used in automatic sense annotation (Section 4.1) and variants of the baseline model based on various pretrained transformers (Section 4.2).

## 2 Experimental Setup

The experiment reflects the proposed WSD-based evaluation protocol: ambiguous words in spontaneous dialog are first disambiguated by dialog systems and then evaluated by humans (or, less interactively, against predefined gold standard data).

### 2.1 Selected Corpus

NEWT-SBCSAE, released by Lưu and Malamud (2020), includes seven *15*-minute extracts of face-to-face casual dialogs from the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000), segmented into *3253* turn-constructional units (TCUs) by Riou (2015) and accompanied by audio files publicly browsable at TalkBank.org. This corpus possesses a rare combination[6] of valuable features:

[4] Including the use of sense relation knowledge encoded in thesauri such as WordNet.

[5] Given that language is continuously changing.

[6] The only existing corpus of its kind I am aware of.

- freely and publicly accessible (in a well-developed XML-based data format)
- carefully curated to include only naturally occurring casual dialogs by a wide variety of people, differing in gender, occupation, social background, and regional origin in comparison with its compact size

The selection of this corpus rests upon the assumption that the corpus can serve as an approximate upper bound of lexical diversity of human–computer spontaneous conversation in the same dialect of English within the evaluation scale of this empirical study. The preference for this corpus over a currently available corpus of human-computer spontaneous conversations is also supported by the fact that the latter may not actually be as representative as claimed (Doğruöz and Skantze, 2021). It is worth noting that the results achieved in this study may not generalize to varieties of American English not present in the corpus, to other regional varieties of English, or to other languages.

### 2.2 Automatic WSD

**Automatic Transcript Preprocessing** After every prosodic token are replaced with "...", each turn-constructional unit (TCU) is tokenized, lemmatized, and part-of speech (POS) tagged by spaCy[7] (v2.3.5)'s small core model for English. Then each ambiguous word is identified as follows:

- its universal POS is in WordNet, i.e. adjective, adverb, noun, proper noun, or verb
- it has more than one WordNet synset (information about the synsets, i.e. sense names and corresponding definitions, is also retrieved)

**SOTA** I use Conia and Navigli (2021) as a SOTA WSD model because it is the back end of AMuSE-WSD[8] (AW), the first end-to-end system that provides a web-based API for downstream tasks to obtain high-quality sense information in 40 languages, including English (Orlando et al., 2021). This model is composed of BERT (large-cased, frozen), a non-linear layer and a linear classifier, and trained on the SemCor corpus (Miller et al., 1994) as well as WordNet glosses and examples with a multi-label classification objective. It achieves 80%-accuracy on the concatenation of all Unified Evaluation Framework datasets for English all-words WSD (Raganato et al., 2017).

The AW API takes as input the text string of each TCU and yields a list of tokens automatically annotated with lemma, POS, and WordNet sense if available. Next, this output sequence is aligned with the spaCy preprocessing output.

**Baseline** The baseline WSD model (cf. Oele and van Noord, 2018) picks the best sense of each ambiguous word (identified in preprocessing) by ranking similarity scores between the contextual embeddings of the word and of the definitions of its WordNet senses, accessed via spacy-wordnet[7]. The contextual embeddings are from DistilBERT (Sanh et al., 2019), accessed via spacy-transformers[7].

### 2.3 Human WSD Judgment

**Task** The models' output was evaluated by two annotators, both Linguistics majors (incl. Formal Semantics) and native speakers of English[9].

For each target word, the annotators saw:

- the WordNet senses assigned to the word by AW and the baseline model[10]
- the list of possible WordNet senses for the word, taking into account its POS

The annotators were asked to decide if:

- AW sense is appropriate (and different from the baseline) – label ***'1'***
- the baseline sense is appropriate (and different from AW) – label ***'2'***
- Both are the same & appropriate – label ***'both'***
- No sense is appropriate and at least one of them has a correct POS – label ***'0'***
- Both senses have incorrect POS and their actual POS are still covered by WordNet – label ***'c'*** (i.e. *'content word but wrong POS'*)
- Both senses have incorrect POS and their actual POS are not covered by WordNet – label ***'f'*** (i.e. *'function word'*)

For ***'0'*** and ***'c'***, the annotators provided the appropriate senses, sometimes from WordNet senses.

The annotation was run in two rounds. In the first round (R.1), both annotators worked on the same dialog so that their inter-annotation agreement (IAA) could be assessed as shown in Table 1(a). The agreement level was substantial (Lan-

[7] Under the MIT License.

[8] Under the CC BY-NC-SA 4.0 License.

[9] From North-Eastern US. They were paid $15–16/hour.

[10] The listing order of these senses are the same for all target words. Consequently, the annotators could recognize that one system is better and treat its prediction as the default for borderline cases, which might slightly inflate the better system's results. On the other hand, this setting reflects real evaluation scenarios in which evaluators are aware of the performance of a specific dialog system throughout their evaluation sessions.

dis and Koch, 1977) and the inter-annotator consistency likely improved after the review of this annotation round and the corresponding revision of annotation guidelines for the final round (R.2), in which the annotators worked on different dialogs.

| | **(a) IAA** | | **(b) Count** | | |
|---|---|---|---|---|---|
| **Tokens** | **Ratio** | **Kappa** | **R.1** | **R.2** | **Total** |
| **all** | 0.750 | 0.660 | 669 | 5681 | 6350 |
| **AW** | 0.741 | 0.641 | 632 | 5366 | 5998 |

Table 1: Statistics of the annotation task.

In Table 1, **all** tokens are the ambiguous words identified in preprocessing; **AW** tokens exclude:

- proper nouns for which AW does not provide WordNet senses
- tokens that AW doesn't tag as adjectives, adverbs, nouns, proper nouns or verbs
- tokens that cannot be aligned with AW outputs

Table 1(b) shows the counts of these types of tokens for each annotation round and in total. The existence of non-AW tokens (5.5% of all tokens in total) demonstrates the challenge of aligning the output of off-the-shelf end-to-end WSD models with the output of the NLP pipeline inherent in a dialog system in real-life situations.

Further annotation details (e.g. data format, platform and examples) can be found in Appendix B.

**Outcome**[11] To facilitate fair comparisons between AW and the baseline WSD model, only AW tokens are considered in the following statistics. In addition, the counts of the first round only cover instances that get the same judgments from both annotators on the aspects the counts concern.

Table 2 shows the various sense judgments, corresponding to the labels listed in Section 2.3.

| | ***'1'*** | ***'2'*** | ***'both'*** | ***'0'*** | ***'c'*** | ***'f'*** | $\sum$ |
|---|---|---|---|---|---|---|---|
| **R. 1** | 200 | 40 | 123 | 94 | 2 | 9 | 468 |
| **R. 2** | 2225 | 440 | 1255 | 1007 | 55 | 384 | 5366 |
| **Total** | 2425 | 480 | 1378 | 1101 | 57 | 393 | 5834 |

Table 2: Counts of the human WSD judgment.

Table 3 shows key statistics as the prerequisite for answering the research questions in Section 1. Table 3(a) shows two groups of sense annotations, based on whether the annotated appropriate sense (unavailable for ***'f'*** cases) is covered by WordNet or not (Section 3). Table 3(b) shows main POS-based groups of sense annotations that are used as gold standard to evaluate automatic WSD effectiveness (Section 4). This data only include cases in which both AW and the baseline senses have correct POS and the appropriate WordNet sense is available.

## 3 WordNet Sense Coverage

WordNet senses cover 96.3% of ambiguous words as shown in Table 3(a). POS-wise, they cover 95.6% adjectives, 98.2% adverbs, 95.7% nouns, 96.6% verbs. Among 200 non-WordNet tokens:

- 1 token is sub-word ("toes" in "Of the different cantos or cantos or whatever toes.")
- 4 tokens are named entities
- 64 tokens are components of multiword expressions or used idiomatically. Handling multiword expressions by feeding phrases instead of tokens into the WordNet search engine would improve the WordNet coverage to 96.7% as more 19 tokens are covered.

So, WordNet coverage for conversations is good.

## 4 Automatic WSD Effectiveness

The gold standard data presented in Table 3(b) covers 1046 lemmas, including 191 adjectives, 80 adverbs, 501 nouns and 274 verbs.

### 4.1 Initial WSD Models

Table 4 shows the performances of AW and the baseline models across POS and in total. The values in ***'both'*** columns illustrate the portion of correct disambiguated senses shared by both models.

AW model performs well on conversational text with the accuracy of 73.7%, though it does not achieve 80% as it did on non-conversational data. In addition, it performs consistently across all POS.

The 36%-level accuracy of the DistilBERT-based baseline model is encouraging, given that the average number of WordNet senses per word token (sense average) is 9.9. Its low performance on verbs can be explained by the high sense average of this POS: 15.5 (versus adjectives – 7.5, adverbs – 4.7, and nouns – 6.3). To improve this model's performance, we can experiment with different ways of manipulating the text containing target words before feeding it into a pretrained transformer.

### 4.2 Experiments with Pretrained Transformers

Table 5 shows the performances of the baseline model, using BERT, XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019), accessed via spacy-transformers. Comparing to the DistilBERT-based

[11] The annotated data is publicly accessible at https://alexluu.flowlu.com/hc/6/271–wsd.

|  | **(a) WordNet sense coverage** |  |  | **(b) Gold standard data** |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | ***yes*** | ***no*** | **∑** | **ADJ** | **ADV** | **NOUN** | **VERB** | **∑** |
| **R.1** | 439 (98.7) | 6 (1.3) | 445 (100) | 68 (16.8) | 61 (15.1) | 149 (36.8) | 127 (31.3) | 445 (100) |
| **R.2** | 4788 (96.1) | 194 (3.9) | 4982 (100) | 538 (11.4) | 755 (16.1) | 1507 (32.1) | 1899 (40.4) | 4699 (100) |
| **Total** | 5227 (96.3) | 200 (3.7) | 5427 (100) | 606 (11.9) | 816 (16.0) | 1656 (32.4) | 2026 (39.7) | 5104 (100) |

Table 3: Statistics (counts and percentages) of the human WSD judgment.

|  | **ADJ** |  |  | **ADV** |  |  | **NOUN** |  |  | **VERB** |  |  | **All** |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | **AW** | **DB** | ***'both'*** | **AW** | **DB** | ***'both'*** | **AW** | **DB** | ***'both'*** | **AW** | **DB** | ***'both'*** | **AW** | **DB** | ***'both'*** |
| **R.1** | 66.2 | 50.0 | 36.8 | 83.6 | 37.7 | 31.1 | 77.9 | 41.6 | 32.2 | 85.8 | 33.1 | 23.6 | 79.3 | 39.8 | 30.1 |
| **R.2** | 74.5 | 42.9 | 32.9 | 76.2 | 37.1 | 30.6 | 76.0 | 45.0 | 33.8 | 69.5 | 25.6 | 17.6 | 73.2 | 35.7 | 26.6 |
| **Total** | 73.6 | 43.7 | 33.3 | 76.7 | 37.1 | 30.6 | 76.2 | 44.7 | 33.7 | 70.7 | 26.1 | 18.0 | 73.7 | 36.0 | 26.9 |

Table 4: Accuracy (%) of initial WSD models (**DB**: DistilBERT).

|  | **ADJ** |  |  | **ADV** |  |  | **NOUN** |  |  | **VERB** |  |  | **All** |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | **B** | **X** | **R** | **B** | **X** | **R** | **B** | **X** | **R** | **B** | **X** | **R** | **B** | **X** | **R** |
| **R.1** | 44.1 | 36.8 | 33.8 | 50.8 | 16.4 | 42.6 | 44.3 | 22.8 | 34.9 | 33.9 | 20.5 | 27.6 | 42.0 | 23.5 | 33.6 |
| **R.2** | 42.2 | 21.7 | 34.2 | 34.6 | 17.9 | 38.1 | 43.5 | 24.0 | 34.6 | 22.7 | 12.5 | 21.7 | 33.5 | 18.1 | 29.9 |
| **Total** | 42.4 | 23.4 | 34.2 | 35.8 | 17.8 | 38.5 | 43.6 | 23.9 | 34.7 | 23.4 | 13.0 | 22.1 | 34.2 | 18.5 | 30.2 |

Table 5: Accuracy (%) of variants of the baseline WSD models (**B**: BERT, **X**: XLNet, **R**: RoBERTa).

model, the performances decrease in the order of [BERT > RoBERTa > XLNet] across POS and in total, except for the case of adverbs in which RoBERTa performs best. XLNet's performance is noticeably low in comparison with the others.

The empirical results show that DistilBERT is the best option for disambiguating WordNet senses of words by ranking similarity scores between contextual embeddings of the words and of the definitions of their senses. DistilBERT is not only effective but also efficient as it is the only simplified version of BERT among the tested transformers.

## 5 Discussion

**Future Work** Next, I will perform a detailed data analysis to gain insights into (1) what the annotators disagreed about, (2) what kinds of errors the WSD models made, and (3) how good incorrect senses are, taking into account the distinction between polysemous and homonymous senses, which is not available in WordNet (Freihat et al., 2016; Habibi et al., 2021; Janz and Maziarz, 2021). These insights will help improve the design of the annotation task and the performance of the WSD models.

I will also study the effect of manipulation of input utterances, by taking into account the linguistic and discourse information about the target words, on the performance of the pretrained transformers. This can shed light on how to create optimal contextual embeddings of ambiguous words for WSD.

**Limitations and Challenges** Exclusively relying on pre-existing sense inventories such as WordNet, the proposed evaluation method would not only miss semantically ambiguous words that do not have multiple senses in these sense inventories, but also inherit their limitations, due to the fact that their senses have different degrees of granularity and cannot keep up with the continuously involving character of natual languages (Mennes and van der Waart van Gulik, 2020; Bevilacqua et al., 2021).

The proposed evaluation method may not easily be adopted by the developers of end-to-end dialog models, the most popular approach to open-domain dialog systems (Huang et al., 2020), as the "black box" nature of these systems does not facilitate human-readable word-level interpretations.

## 6 Conclusion

This work proposes WSD, an established NLP task, as a required component of a valid and reliable human evaluation framework for mutual understanding in human–computer spontaneous conversation. The conducted experiments demonstrate the practicality of this proposal for English. To sufficiently evaluate human–computer mutual understanding, I envision that the WSD component will be necessarily coupled with a reasoning judgment component in which human evaluators assess the appropriateness of conversation moves made by a dialog system, including clarifying and adjusting their interpretations, based on the disambiguated word senses in those moves. This setting will help human evaluation become more grounded and therefore more objective than the current common practices, in which human evaluators are asked to rate system responses using vaguely defined criteria and inconsistent numeric scales (Finch and Choi, 2020).

## Acknowledgements

The annotation work of this study was funded by a PhD Research Award from the Graduate School of Arts and Sciences, Brandeis University. I am extremely grateful to my annotators, Josh Broderick Phillips and Tali Tukachinsky, for their diligence and professionalism. My deepest gratitude goes to Sophia A. Malamud, who exhaustively discussed every aspect of this study with me. Finally, I would like to thank Nianwen Xue, the anonymous ARR reviewers of the January 2022 deadline and the organizers of HumEval at ACL 2022 for their detailed, constructive and actionable feedback.

## References

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, 32(8):827–877. Publisher: Routledge _eprint: https://doi.org/10.1080/09588221.2018.1535508.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, Cambridge.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

A. Seza Doğruöz and Gabriel Skantze. 2021. How "open" are the conversations with open-domain chatbots? a proposal for speech event based evaluation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 392–402, Singapore and Online. Association for Computational Linguistics.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Christiane Fellbaum. 2010. WordNet. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands, Dordrecht.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2016. A taxonomic classification of WordNet polysemy types. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 106–114, Bucharest, Romania. Global Wordnet Association.

Amir Ahmad Habibi, Bradley Hauer, and Grzegorz Kondrak. 2021. Homonymy and polysemy detection with multilingual information. In *Proceedings of the 11th Global Wordnet Conference*, pages 26–35, University of South Africa (UNISA). Global Wordnet Association.

Mika Hämäläinen and Khalid Alnajjar. 2021. The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74, Online. Association for Computational Linguistics.

Hendrik Heuer and Daniel Buschek. 2021. Methods for the design and evaluation of HCI+NLP systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–33, Online. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Transactions on Information Systems*, 38(3):21:1–21:32.

Arkadiusz Janz and Marek Maziarz. 2021. Discriminating homonymy from polysemy in wordnets: English, Spanish and Polish nouns. In *Proceedings of the 11th Global Wordnet Conference*, pages 53–62, University of South Africa (UNISA). Global Wordnet Association.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Alex Lưu and Sophia A. Malamud. 2020. Annotating coherence relations for studying topic transitions in social talk. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 174–179, Barcelona, Spain. Association for Computational Linguistics.

Julie Mennes and Stephan van der Waart van Gulik. 2020. A critical analysis and explication of word sense disambiguation as approached by natural language processing. *Lingua*, 243:102896.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Dieke Oele and Gertjan van Noord. 2018. Simple embedding-based word sense disambiguation. In *Proceedings of the 9th Global Wordnet Conference*, pages 259–265, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 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. Association for Computational Linguistics.

Marine Riou. 2015. *The Grammar of Topic Transition in American English Conversation. Topic Transition Design and Management in Typical and Atypical Conversations (Schizophrenia)*. Ph.D. thesis, Université Sorbonne Paris Cité.

Stephen Roller. ParlAI tutorial (accessed on 12/12/2021).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, Vancouver, Canada.

Lucia Specia. 2021. Disagreement in human evaluation: Blame the task not the annotators. Invited talk at the Workshop on Human Evaluation of NLP systems (HumEval).

Maarten van Gompel, Ko van der Sloot, Martin Reynaert, and Antal van den Bosch. 2017. *FoLiA in Practice: The Infrastructure of a Linguistic Annotation Format*, pages 71–82. Ubiquity Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

# A An Example of Divergence in Human Interpretation

12 native speakers of American English (2 PhD, 9 master's, 1 senior undergraduate) in a linguistic course are asked to give their interpretation of entities available in the following excerpt of dialogue between Jim and Michael, adapted from this publicly accessible recording (10'32"–11'04"):

> **Jim**: *So much of today's technology is soulless and has nothing to do with peace. It has to do with chewing up the human experience and turning it into some kind of consumer need.*
>
> **Michael**: *Did you ever get into Tesla?*
>
> **Jim**: *Just ever so peripherally.*
>
> **Michael**: *He had a lot of real wacky ideas on big levels. He wanted a world power system, that you could tap into the air basically, and get power anywhere on earth.*

The interpretation results for the token "Tesla" and the corresponding pronouns "he" is presented in Table 6.

| "Tesla" | "he" | Count |
|---|---|---|
| Nicola Tesla | Nicola Tesla | 6 |
| Nicola Tesla's body of work | Nicola Tesla | 4 |
| Tesla, Inc. | Nicola Tesla | 1 |
| Tesla, Inc. | Elon Musk, CEO of Tesla, Inc. | 1 |

Table 6: Divergence in human interpretation.

# B Annotation in Practice

## B.1 Annotation Data Format and Platform

The annotation files are stored in the XML-based FoLiA format[12], which accommodates multiple

[12] An open file format, whose specification and documentation are generated by open source code under GNU General Public License version 3.0.

linguistic annotation types with arbitrary tagsets, and annotated with FLAT[13], FoLiA's web-based annotation tool whose user-interface can show different linguistic annotation layers at the same time (van Gompel et al., 2017).

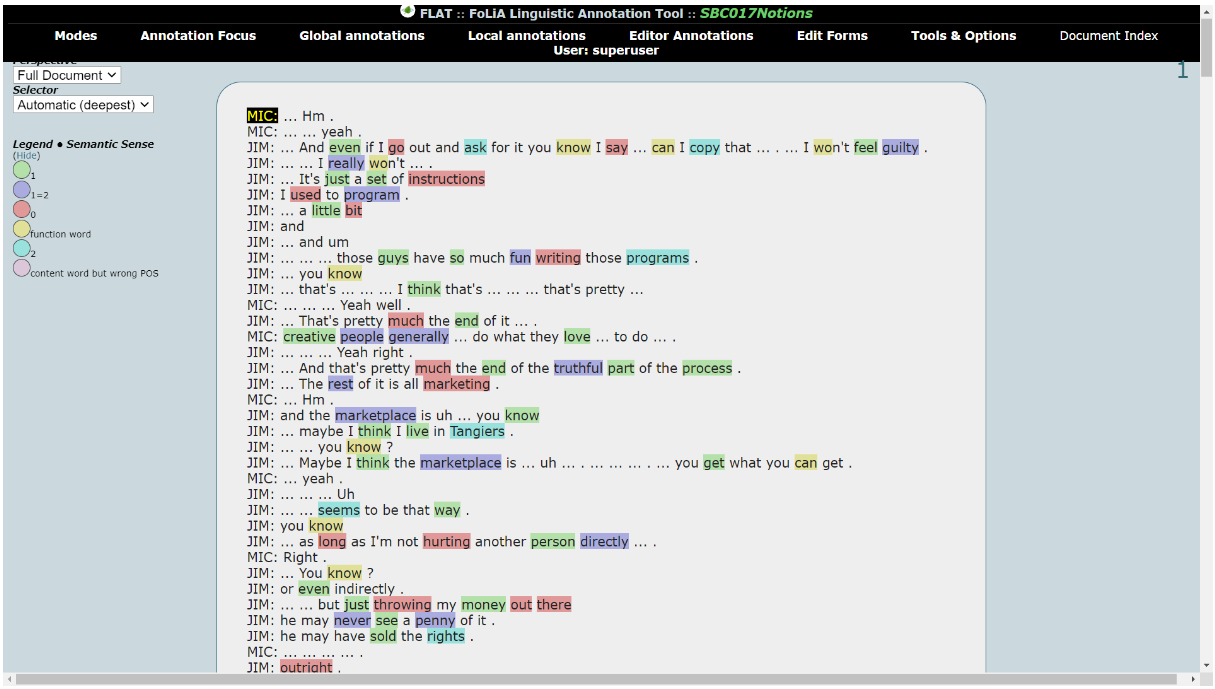
### B.2 Annotation Examples

Figures 2–4 display an annotation file opened on FLAT. The ambiguous words are highlighted in different colors, corresponding to the annotation labels mentioned in Section 2.3, so that the annotators can navigate them quickly.

Figure 3 shows that when a word token such as "guilty" is hovered over, it is highlighted in black while its text turns yellow, and all of its annotation information are displayed in a pop-up box.

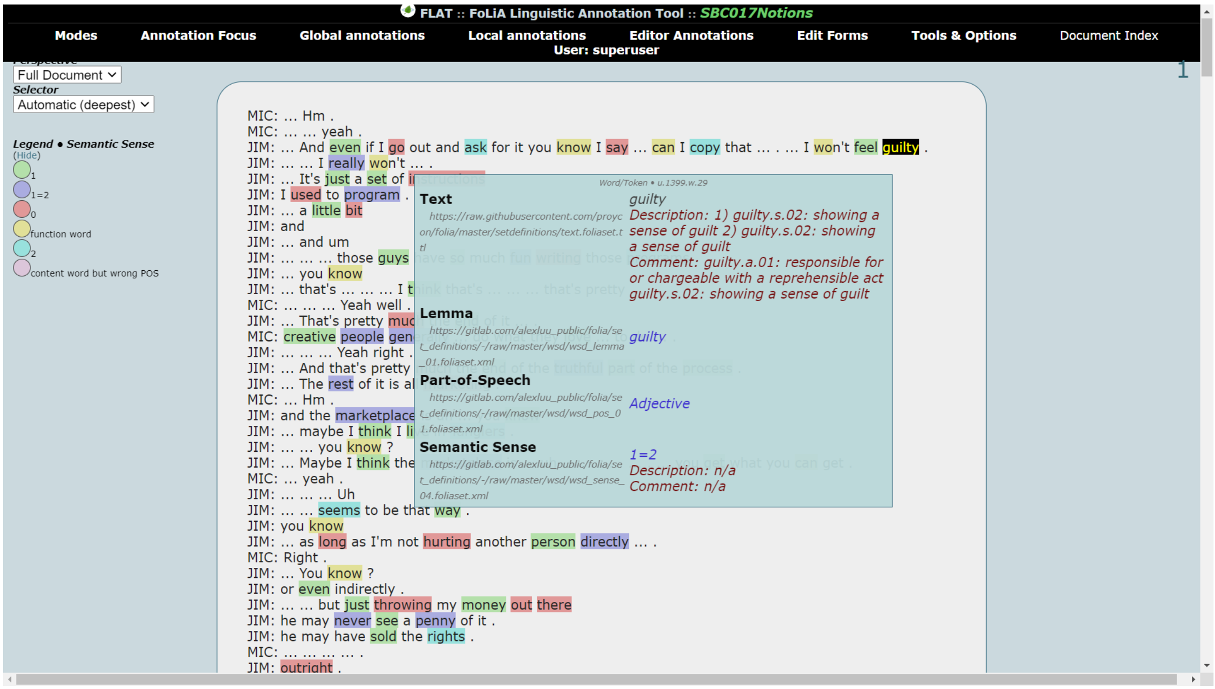
Figure 4 shows that when "guilty" is clicked, it is highlighted in yellow, and its annotation layers become editable in the **Annotation Editor**.

[13] Under GNU General Public License version 3.0.



MIC: … Hm .
MIC: … … yeah .
JIM: … And even if I go out and ask for it you know I say … can I copy that … . … I won't feel guilty .
JIM: … … I really won't … .
JIM: … It's just a set of instructions
JIM: I used to program .
JIM: … a little bit
JIM: and
JIM: … and um
JIM: … … … those guys have so much fun writing those programs .
JIM: … you know
JIM: … that's … … … I think that's … … … that's pretty …
MIC: … … … Yeah well .
JIM: … That's pretty much the end of it … .
MIC: creative people generally … do what they love … to do … .
JIM: … … … Yeah right .
JIM: … And that's pretty much the end of the truthful part of the process .
JIM: … The rest of it is all marketing .
MIC: … Hm .
JIM: and the marketplace is uh … you know
JIM: … maybe I think I live in Tangiers .
JIM: … … you know ?
JIM: … Maybe I think the marketplace is … uh … . … … … . … you get what you can get .
MIC: … yeah .
JIM: … … … Uh
JIM: … … seems to be that way .
JIM: you know
JIM: … as long as I'm not hurting another person directly … .
MIC: Right .
JIM: … You know ?
JIM: or even indirectly .
JIM: … … but just throwing my money out there
JIM: he may never see a penny of it .
JIM: he may have sold the rights .
MIC: … … … … .
JIM: outright .

Figure 2: Annotation interface on FLAT.



MIC: … Hm .
MIC: … … yeah .
JIM: … And even if I go out and ask for it you know I say … can I copy that … . … I won't feel guilty .
JIM: … … I really won't … .
JIM: … It's just a set of i
JIM: I used to program .
JIM: … a little bit
JIM: and
JIM: … and um
JIM: … … … those guys
JIM: … you know
JIM: … that's … … … I t
MIC: … … … Yeah well .
JIM: … That's pretty muc
MIC: creative people gen
JIM: … … … Yeah right .
JIM: … And that's pretty
JIM: … The rest of it is al
MIC: … Hm .
JIM: and the marketplace
JIM: … maybe I think I li
JIM: … … you know ?
JIM: … Maybe I think the
MIC: … yeah .
JIM: … … … Uh
JIM: … … seems to be that way .
JIM: you know
JIM: … as long as I'm not hurting another person directly … .
MIC: Right .
JIM: … You know ?
JIM: or even indirectly .
JIM: … … but just throwing my money out there
JIM: he may never see a penny of it .
JIM: he may have sold the rights .
MIC: … … … … .
JIM: outright .

*Word/Token • u.1399.w.29*

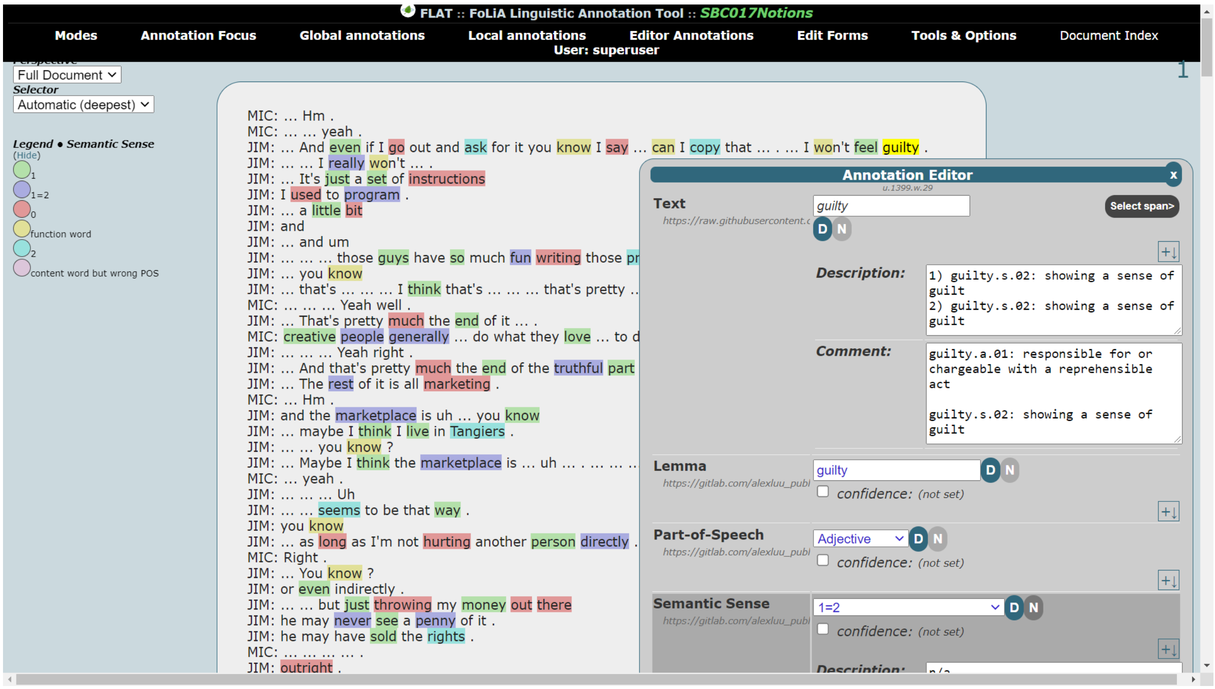
**Text** *https://raw.githubusercontent.com/proycon/folia/master/setdefinitions/text.foliaset.ttl*
*guilty*
*Description: 1) guilty.s.02: showing a sense of guilt 2) guilty.s.02: showing a sense of guilt*
*Comment: guilty.a.01: responsible for or chargeable with a reprehensible act guilty.s.02: showing a sense of guilt*

**Lemma** *https://gitlab.com/alexluu_public/folia/set_definitions/-/raw/master/wsd/wsd_lemma_01.foliaset.xml*
*guilty*

**Part-of-Speech** *https://gitlab.com/alexluu_public/folia/set_definitions/-/raw/master/wsd/wsd_pos_01.foliaset.xml*
*Adjective*

**Semantic Sense** *https://gitlab.com/alexluu_public/folia/set_definitions/-/raw/master/wsd/wsd_sense_04.foliaset.xml*
*1=2*
*Description: n/a*
*Comment: n/a*

Figure 3: Quick access to the annotation information of a token.



Perspective: Full Document

Selector: Automatic (deepest)

Legend • Semantic Sense

- 🟢 1
- 🟣 1=2
- 🔴 0
- 🟡 function word
- 🔵 2
- ○ content word but wrong POS

```
MIC: … Hm .
MIC: … … yeah .
JIM: … And even if I go out and ask for it you know I say … can I copy that … . … I won't feel guilty .
JIM: … … I really won't … .
JIM: … It's just a set of instructions
JIM: I used to program .
JIM: … a little bit
JIM: and
JIM: … and um
JIM: … … … those guys have so much fun writing those pr
JIM: … you know
JIM: … that's … … … I think that's … … … that's pretty ..
MIC: … … … Yeah well .
JIM: … That's pretty much the end of it … .
MIC: creative people generally … do what they love … to d
JIM: … … … Yeah right .
JIM: … And that's pretty much the end of the truthful part
JIM: … The rest of it is all marketing .
MIC: … Hm .
JIM: and the marketplace is uh … you know
JIM: … maybe I think I live in Tangiers .
JIM: … … you know ?
JIM: … Maybe I think the marketplace is … uh … . … … ..
MIC: … yeah .
JIM: … … … Uh
JIM: … … seems to be that way .
JIM: you know
JIM: … as long as I'm not hurting another person directly .
MIC: Right .
JIM: … You know ?
JIM: or even indirectly .
JIM: … … but just throwing my money out there
JIM: he may never see a penny of it .
JIM: he may have sold the rights .
MIC: … … … … .
JIM: outright .
```

Annotation Editor

u.1399.w.29

**Text** https://raw.githubusercontent.c — guilty

*Description:*

```
1) guilty.s.02: showing a sense of
guilt
2) guilty.s.02: showing a sense of
guilt
```

*Comment:*

```
guilty.a.01: responsible for or
chargeable with a reprehensible
act

guilty.s.02: showing a sense of
guilt
```

**Lemma** https://gitlab.com/alexluu_publ — guilty
□ *confidence:* *(not set)*

**Part-of-Speech** https://gitlab.com/alexluu_publ — Adjective
□ *confidence:* *(not set)*

**Semantic Sense** https://gitlab.com/alexluu_publ — 1=2
□ *confidence:* *(not set)*

*Description:* n/a

Figure 4: **Annotation Editor** for a token.