Potentially Problematic Word Usages and How to Detect Them: A Survey

Aina Garí Soler[1] **Matthieu Labeau**[2] **Chloé Clavel**[1]
[1]INRIA, Paris, [2]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.gari-soler,chloe.clavel}@inria.fr, matthieu.labeau@telecom-paris.fr

Abstract

We introduce and frame the concept of potentially problematic word usages (PPWUs): word occurrences that are likely to cause communication breakdowns of a semantic nature. While much research has been devoted to lexical complexity, ambiguity, vagueness and related issues, no work has attempted to fully capture the intricate nature of PPWUs. We review linguistic factors, datasets and metrics that can be helpful for PPWU detection. We also discuss challenges to their study, such as their complexity and subjectivity, and highlight the need for future work on this phenomenon.

1 Introduction

Language is a powerful communication tool allowing us to exchange complex messages, but information is not always conveyed successfully. Miscommunication can be due to multiple factors, both linguistic and non-linguistic (e.g., environmental or psychological causes, such as a noisy background or the listener's lack of attention). In this paper, we focus on cases where a specific use of a word can give rise to a misunderstanding or an objection related to its meaning.

We define a **potentially problematic word usage (PPWU)** as an instance of a word[1] in context which is likely to cause some type of miscommunication (misunderstanding or non-understanding) or disagreement of a semantic nature.[2] For example, in "I enjoy working on my car", it is not clear whether *working* means polishing it as a hobby or repairing it (see Table 1). In practice, PPWUs should be determined with a specific target population in mind. In this paper, we discuss intrinsic factors of words and their contexts which are problematic both for specific communities but also for the general public (e.g., words with a false friend in a language may only risk being misunderstood or misused by speakers of that language, but a word in an underconstrained context can be problematic for anyone). We say, of a specific word usage, that it is an (actual) problematic word usage (PWU) when there is evidence that it has been misunderstood or disagreed upon by someone (e.g., when someone has signalled it in an interaction, asking "what do you mean by ...?" or similar questions (Noble et al., 2021)).

The detection of PPWUs has several applications. It can be useful for text simplification and readability assessment and can also have uses in applications related to language learning in general, such as aiding in choosing the right learning materials to adapt them to a student's level, or designing exercises that target specific types of PPWUs. If integrated with a conversational system, PPWUs could either be actively avoided in production, or trigger directed clarification requests to the user. Ensuring clear language use contributes to decrease misunderstandings, which can have negative psychological and physiological effects (Crockett et al., 2022). As part of a writing assistance tool, it could help identify words to be replaced to improve clarity. It could also be used to detect lexical errors in translated text.

Numerous studies focus on specific types or causes of PPWUs, but a unifying perspective encompassing them in their full complexity is missing. In this paper, we introduce and frame the notion of PPWU bringing together insights and points of view from research on different domains (e.g, psycholinguistics, computational linguistics, cognitive science, NLP) and with different goals, providing a foundation and framework for future research on PPWU detection. To this end, we compile and outline multiple reasons why a specific word or word usage is likely to cause miscommunication or be disagreed upon, thereby proposing a first character-

[1]We consider open-class lexemes or lexical items in general, but refer to them as "words" for simplicity.

[2]We exclude cases of unclear referents of referring expressions such as pronouns or proper nouns.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 441–463

ization of PPWUs. We also present existing computational methods, resources and datasets that can be helpful for the study and detection of PPWUs, identifying areas where more work and knowledge is necessary; and include a discussion on various considerations to make when conducting research on PPWUs.

To compile the bibliography for this survey, we searched for relevant literature on Google Scholar using keywords such as "polysemy detection," "false friends," "lexical complexity," "text readability," or "lexical errors;" and expanded the search by examining papers citing, or cited by, the returned papers.

2 Factors contributing to PPWUs

In this section we provide a non-exhaustive[3] list of linguistic properties and phenomena linked to PPWUs. We distinguish between factors tied exclusively to word identity (Section 2.1) and factors linked to the linguistic context where a word is used (Section 2.2).[4] In the Appendix, we provide examples (Table 1) and a diagram summarizing the content of this section (Figure 1).

2.1 Context-independent factors

Some words are inherently difficult to understand, regardless of the context they appear in. This is often referred to as **lexical complexity** (North et al., 2023). Here, we outline word characteristics that may be *reasons* for a word not to be understood or for its meaning to be disagreed upon.[5] Note that the properties presented here are not absolute determinants of PPWUs; they often interact with each other as well as with other linguistic variables.[6] While we have aimed to define distinct categories where possible, they are not mutually exclusive and can co-occur in real-world usage.

[3]Our goal is to propose a categorization of PPWU causes linked to relevant NLP tasks, but there are other idiosyncratic causes of PPWUs that are difficult to categorize.

[4]This distinction bears some parallels to the notions of "meaning potential" and "situated meaning" (Myrendal, 2019).

[5]Some of the notions presented are used as features in lexical complexity detection. We exclude factors that, while correlated with complexity, are not likely to cause miscommunication on their own (e.g., word length and syllable count (Desai et al., 2021)).

[6]For example, a low-frequency word that is morphologically transparent (*cardiomyopathy*) may be easier to understand than another rare word with an opaque morphology (*gybe*).

2.1.1 Word properties

Lexical ambiguity. Words that have multiple senses (homonyms and polysemous words) are more likely to be misunderstood, even if the context is enough for disambiguation, because the audience might not be familiar with all of their senses, especially if some of them are not very frequent. The number of polysemous words and unique senses in a text has been found to correlate with its readability (Danilov et al., 2023), and the readers' knowledge of the multiple senses of a word correlates with reading comprehension (Kenneth Logan and Kieffer, 2017; Booton et al., 2022). Martínez Alonso et al. (2015) found that, in a sense annotation task, words with a higher number of senses and sense entropy tend to display higher disagreement. Words that have undergone lexical semantic changes (LSC) can be particularly difficult: readers or listeners are less likely to know novel senses of a word if these are recent, or, if reading a text from a different time period, a word may have changed in a way that the reader may not be aware of (e.g., *gay* used to mean *light-hearted, cheerful*). Approaches to quantifying a word's number of senses (**polysemy detection**, Garí Soler and Apidianaki (2021a)) and whether, how and when a word has changed meaning (**LSC detection**, Schlechtweg et al. (2020); Montariol et al. (2021)) can be helpful in finding words that are more likely to be misunderstood.

Word frequency is one of the most useful features to estimate lexical complexity (Specia et al., 2012; Wilkens et al., 2014; Garí et al., 2018). This is not surprising, as rare words are less likely to be part of a speaker's vocabulary, because they may have had less or no exposure to them. Acronyms, if not commonly used and not introduced properly in a text, are also a common source of confusion. Numerous studies have shown the relationship between word frequency and reading comprehension (Marks et al., 1974; Freebody and Anderson, 1983; Nouri and Zerhouni, 2018), confirming that rarer words are more prone to limit comprehension than more frequent words. Word frequency lists exist for multiple languages (Speer, 2022). Their source is very important: the nature, register, variety and size of the corpus, among other factors, may determine the usefulness of word frequency estimations as features for lexical complexity prediction (Wilkens et al., 2014).

Neologisms are newly created words recently introduced into a language, and which have a certain degree of acceptance in a linguistic community (e.g., *rizz*). As such, speakers (especially non-native ones (Charteris-Black, 1998)) are less likely to know them. An additional difficulty of neologisms is that they are often driven by technological advancements, so certain speakers may be unfamiliar not only with a new word form but also with the concept it expresses. At the same time, neologisms are often coined from existing morphemes (e.g., *mansplaining*) or from other languages, which may make their interpretation easier than that of neologisms created *ex nihilo* (e.g., *cromulent*) (Lehrer, 2003). Although less common, grammatical neologisms (existing words used with a new part of speech (PoS)) can also be problematic. Approaches to **neologism detection** (Janssen, 2012; Falk et al., 2014; Klosa and Lüngen, 2018) can be useful for detecting PPWUs. Finally, semantic neologisms (the appearance of new senses) are a form of lexical semantic change which can be addressed with the task of **novel sense detection** (Section 2.2).

Lexical variation refers to differences in lexical choice due to factors such as age, profession and social class (sociolects) or geographical location (dialectal variation). This includes slang, jargon and dialect-specific words which may only be produced and understood by certain communities of speakers. These differences can be particularly problematic in heterogeneous conversations (with people from, e.g., different age groups or cultures). Work on detecting **synchronic lexical differences** (Gonen et al., 2020; Yin et al., 2018; Schlechtweg et al., 2019; Garimella et al., 2016) and **dialect lexicon induction** (Scherrer, 2007; Artemova and Plank, 2023) could help identify words that are distinctive of certain groups of speakers.

Idioms are phrases with a non-compositional meaning; i.e., their meaning cannot be inferred from their parts (e.g., *kick the bucket*). As such, they need to be learned as autonomous lexical items. It is well known that idioms are particularly challenging for second language (L2) learners (Schraw et al., 1988; Alhaysony, 2017), except when a similar idiom exists in their native language (Irujo, 1986). Regardless of the frequency of an idiom, familiarity with the words that compose it can give a false perception of understanding. In fact, learners tend to overestimate their comprehension of idioms made up of high frequency words (Martinez and Murphy, 2011; Park and Chon, 2019); and when analyzing a sample of English idioms, Libben and Titone (2008) found that the frequency of verbs in idioms was negatively correlated with the predictability of their meaning (see Section 5 for more considerations on undetected misunderstanding). Relevant NLP tasks include **idiomatic expression identification** (Zeng and Bhat, 2021), the distinction between **literal and non-literal** idiom usages (Li and Sporleder, 2010), and **compositionality detection** (Cordeiro et al., 2019).

2.1.2 Concept-related properties

Complex meaning. Words designing complex concepts may be misunderstood because of a lack of or an incomplete knowledge of the reality being described (e.g., *enthalpy*). While estimating the complexity of a concept is a hard and subjective task, as it is not a well-established notion, **automatic term extraction** can be a good proxy to find words designating advanced or technical concepts (Hätty et al., 2020; Rigouts Terryn et al., 2020).

Vagueness and generality. Some words have an inherently vague meaning, i.e., a meaning that lacks precision (Van Rooij, 2011). This often concerns gradable adjectives, both relative (such as *big* or *tall*) and absolute (e.g., *bald*, *flat*), but it can also be found in words of other PoS (e.g., *heap*, *idiot*). These words describe qualities for which it is hard to draw a line and which can have multiple interpretations. Pezzelle and Fernández (2023) show that when faced when unclear gradable adjectives, speakers can increase their alignment with explicit interaction about word meaning. Adjectives may inherently be more problematic than other PoS because their semantic contribution to a noun is highly variable and combination-specific (Boleda et al., 2013).

Words that are very general, or high in a hypernymy hierarchy (e.g., *thing* or *do*), may not be specific enough and require clarification or more details. A mismatch between the provided level of specificity and the level required by the communicative situation, i.e., flouting Grice's maximum of quantity (Grice, 1975), can also generate confusion (Cruse, 1977).

Scalar adjective identification (Garí Soler and Apidianaki, 2021b) can be useful to identify potentially vague adjectives and distinguishing them from relational ones (e.g., *wooden*). The VAGO system (Icard et al., 2022) for measuring vague-

ness in texts relies on a database with lexical vagueness information (Atemezing et al., 2022). Lexicons and ontologies such as WordNet (Fellbaum, 1998) can serve as references for generality, and work on **semantic content quantification** (Herbelot and Ganesalingam, 2013; Santus et al., 2014) and more specifically on **hypernymy detection** (Shwartz et al., 2016; Cho et al., 2020) can help identify words with a general meaning.

Connotatively loaded terms. Words related to or referring to controversial topics are also prone to cause disagreement and misunderstanding: speakers may have different mental representations of these words, which are affected by their own opinions on the topic. For example, speakers may detect a disagreement on what constitutes or qualifies as *sexism* or *abuse* and signal this in a conversation (Myrendal, 2019). Work on **controversial topic detection** (Choi et al., 2010; Garimella et al., 2018) as well as on **lexico-semantic alignment** in, or outside of, conversations (Garí Soler et al., 2022, 2023) can potentially help detect this kind of words.

2.1.3 Cross-linguistic influence

In this section we describe cross-linguistic factors that may result in PPWUs when the situation involves L2 learners or bilingual individuals in a particular language pair.

False friends and partial cognates. Two words in two different languages are said to be false friends if they sound or look similar but have different meanings (e.g., *embarrassed* and Spanish *embarazada*, which means *pregnant*). A related concept is that of partial cognates, where the two similar words share some, but not all, meanings (see Table 1). These words can easily give rise to confusion if they have not been learned properly. Work on **false friend** or **partial cognate detection** (Inkpen et al., 2005; Mitkov et al., 2007; Ljubešić and Fišer, 2013; Palmero Aprosio et al., 2020; Lefever et al., 2020; Kanojia et al., 2021) can help identify words that may be problematic, both in terms of production (Raušer, 2017) and comprehension (Mattheoudakis and Patsala, 2007), when a speaker of a particular language is involved.

Cross-linguistic inequivalence. Every language offers different conceptualizations of the world and maps words to referents in different ways (Pavlenko, 2009). There is rarely a 100% translation equivalence between words in two languages, and often words that are almost equivalent differ in specific nuances that are hard to notice and to master, for language learners (Shalaby et al., 2009) but also bilingual speakers (Ameel et al., 2005). These differences can be quite notorious[7] or very subtle, such as what is reflected in the categorization and naming of similar objects (Malt et al., 2003; Pavlenko and Malt, 2011). This is sometimes referred to as "cross-linguistic near-synonymy" (Gries et al., 2020) and is studied in the fields of contrastive lexicology and lexico-semantic typology (Schapper and Koptjevskaja-Tamm, 2022). In other cases, for cultural or historical reasons, some terms in a language may not exist at all in another language, because they designate realities that do not exist in the other culture (e.g., Russian форточка (fortochka), a specific kind of small window for ventilation). These differences can result in unintended cross-linguistic transfer during production and comprehension by multilingual individuals (Jarvis, 2011), which can lead to misunderstandings. An NLP task that could assist in automatically finding cross-linguistic near-synonyms is **bilingual lexicon induction** (Irvine and Callison-Burch, 2017).

Words from another language. Sometimes multilingual speakers interject words from other languages which may be unknown to their interlocutors or readers. **Code switching detection** (Samih et al., 2016; Kevers, 2022) and **language identification** algorithms can detect these usages, particularly when they are tailored to the word level (Solorio et al., 2014; Rijhwani et al., 2017; Ansari et al., 2021).

2.2 Context-dependent factors

A word usage may also be misunderstood or disagreed upon because of the characteristics of the context in which it is used, even if the word itself is not usually problematic.[8] See Table 1 in the Appendix for examples.

Contextual underspecification. While many words have multiple senses, context is often enough for disambiguation, and in practice humans do not have a problem understanding many polysemous word usages (Piantadosi et al., 2012). When this

[7] *meat* is often translated as мясо (myaso) in Russian, but in its everyday use, мясо does not include poultry.

[8] It is, however, not possible to fully disentangle context from word identity: whether a word's context is problematic strongly depends on the word itself.

is not the case, however, ambiguity may require clarification. This happens with underconstrained contexts which do not provide enough information to establish the sense of a word, and which are a common reason of disagreement in word sense annotation, especially when allowing only the annotation of a single sense (Jurgens, 2014). There is not much work on identifying such ambiguities at the lexical level (Liu et al., 2023). Quantifying the **informativeness** of a context with respect to a specific target word could be helpful (Montariol et al., 2019; Schick and Schütze, 2019), but existing approaches assume that the target word is unknown.

Novel senses and metaphors. A speaker may use a word in a new sense that it has only recently acquired, or propose a creative or figurative use of an existing word. The latter are called novel metaphors (i.e., metaphors that do not rely on established conceptual mappings such as LOVE IS A JOURNEY (Lakoff, 1993)), and are harder to understand than conventional metaphors (Lai et al., 2009; Horvat et al., 2022). There is work on **unknown** or **novel word sense detection** (Erk, 2006; Cook et al., 2014), as well as **novel metaphor detection** (Schulder and Hovy, 2014; Haagsma and Bjerva, 2016; Reimann and Scheffler, 2024). One challenge is distinguishing these creative uses of language from errors, discussed below.

Lexical errors. An existing word may be inappropriately used instead of a correct alternative for multiple reasons. In text, a word can be confused with a homophone or another similar-sounding term, causing malapropisms (e.g., *insurance* for *assurance*). Sometimes a speaker may mix words that have a similar or related meaning (*broth* and *stock* or *trip* and *journey*) (Shalaby et al., 2009), or may use a different word because they can't come up with the correct one. Lexical errors can also be due to misspellings which result in another existing word (so-called "real-word errors," like *angel* for *angle*) (Azmi et al., 2019), or be caused by automatic correction tools. Of course, native and non-native speakers may make different mistakes. There is abundant work studying the types of lexical errors encountered in essays written by non-native speakers (Hemchua et al., 2006; Saud, 2018), but not so much about the kinds of mistakes made by native speakers other than malapropisms (Hirst et al., 1998), presumably because they are much rarer and harder to detect. Not all mistakes are equally confusing, however. While exposed to linguistic input, humans develop expectations about what is going to be said, and surface form similarity may facilitate understanding. There is also evidence that native speakers adapt their expectations when faced with non-native speech, and are more likely to find interpretations for implausible statements (Lev-Ari, 2015; Gibson et al., 2017).

Studies targeting anomalies at the lexical level aim at detecting text obfuscation (deliberate word substitutions to encrypt a message (Fong et al., 2008), for example to bypass censorship (Ji and Knight, 2018)); at identifying **real-word misspellings** (Samanta and Chaudhuri, 2013; Bravo-Candel et al., 2021) or detecting **miscollocations** (Wanner et al., 2013).

Rare senses. As discussed in Section 2.1.1, polysemous words used in a rare sense may be problematic. Precisely due to their low frequency, Word Sense Disambiguation (WSD) and Neural Machine Translation (NMT) systems also struggle with them (Campolungo et al., 2022a). Efforts toward **identifying rare word senses** (McCarthy et al., 2004) and correctly **disambiguating** them in context (Barba et al., 2021; Hangya et al., 2021; Campolungo et al., 2022b) can be helpful to find this kind of PPWUs.

Conversational Maxims Flouting. Going against conversational maxims (Grice, 1975) can cause confusion for the listener, especially the maxim of quality, with contradictory statements, jokes, information that goes against common sense, or **vandalism** in, for example, collaborative text editing (Adler et al., 2011). Approaches for **semantic plausibility** (Ko et al., 2019) and **joke detection** (Baranov et al., 2023) can be relevant for these usages, but they are typically designed to work at the sentence or text level.

3 Data

In this section we describe the kinds of data that can be used to investigate PPWUs or to train systems to automatically detect them. Our focus is on datasets of actual PWUs where there is evidence of speaker differences in word meaning. For datasets dedicated to the linguistic factors described in Section 2, refer to Table 2 in the Appendix. The most obvious clues come from real, spontaneous interactions where a speaker explicitly signals a problem with a word used in a conversation (Section 3.1).

Non-dialogical types of data exposing word usage or comprehension differences between speakers are presented in Sections 3.2 and 3.3.

3.1 Dialog

When miscommunication happens in dialog, if it is detected, it can be addressed by means of repair strategies, such as rephrasing or asking for clarification (Purver et al., 2003; Pickering and Garrod, 2004). While some datasets on clarification questions (Xu et al., 2019; Aliannejadi et al., 2019; Kumar and Black, 2020) and repair strategies exist (Rasenberg et al., 2022), they are rarely dedicated to problems with word meaning.

The most relevant work is on **word meaning negotiation** (WMN) in dialog, both in language learning settings (Varonis and Gass, 1985) as well as in online discussions (Myrendal, 2019). WMN can be understood as a case of conversational repair targeting the meaning of a specific word occurrence. WMN instances typically have three parts: a trigger (the PWU), an indicator (the turn signaling a clarification request or an objection), and the meaning negotiation, which is a metalinguistic discussion where speakers explicitly discuss the meaning of the trigger word (Varonis and Gass, 1985; Myrendal, 2015). Myrendal (2015) distinguishes two types of WMN: those which arise from an incomplete understanding and those caused by a disagreement about how someone has used a word. Currently, only one dataset with WMN annotations exists: the NeWMe corpus (Garí Soler et al., 2025a). NeWMe contains over 600 WMNs and related phenomena, coming from both oral and online conversations and involving a PWU. This kind of data can shed light on more aspects and characteristics of (P)PWUs. For example, in an analysis of triggers in NeWMe, Noble et al. (2025) found that disagreement problems tend to involve abstract terms, while understanding issues are more often linked to concrete terms. However, the data remains scarce; given the variety of PPWUs, more annotations covering a wider range of conversational situations are needed for their study.

3.2 Monolog

In monolog-like text, due to the absence of feedback, it is typically harder to anticipate what parts of the discourse may be unclear. In fact, speakers tend to overestimate their interlocutor's understanding (Keysar and Henly, 2002). Collaborative writing environments like Wikipedia, where texts are revised and edited by multiple authors, provide however a useful testbed to investigate problematic language use, by highlighting what can be improved in the original text. The wikiHowToImprove dataset (Anthonio et al., 2020) contains 2.7 million sentences with their revisions. The advantage of wikiHow compared to other sources of **revision histories** like Wikipedia (Faruqui et al., 2018) or news (Spangher et al., 2022) is that, in the former, modifications are more likely to be linguistically motivated instead of updating factual knowledge or providing additional information.

In this kind of interactive setting, PPWUs can be studied through word replacements. For example, Anthonio and Roth (2020) investigate noun substitutions in wikiHowToImprove. The fact that a word is replaced, however, is not evidence of it being a PWU: modifications are not always of a semantic nature (e.g., misspelling corrections), and only about 70% of revised versions in the dataset were judged to be improvements. Indeed, revisions may introduce vandalism, serve to just improve general textual coherence, or simply reflect the editor's personal lexical choice preferences (e.g., replacing *start* with *begin*). A dataset of word replacements annotated to indicate whether they constitute a semantic improvement is currently missing. Other kinds of modifications that can provide interesting data for the study of PPWUs are specific kinds of word insertions that clarify a word usage, such as completions of underspecified noun phrases (*tank* → *goldfish tank*) (Roth et al., 2022).

Finally, another useful kind of data could be obtained by directly asking annotators to mark, in a text, the word usages that they do not understand.

3.3 Signals from annotator disagreement

We can also find evidence of PWUs in cases of low inter-annotator agreement in semantic annotations at the word level. For example, in datasets where multiple people annotated senses (Jurgens and Klapaftis, 2013) or word usage similarity (Erk et al., 2009, 2013; Schlechtweg et al., 2021, 2025). Disagreement can be due to multiple reasons, such as inadequacy of the labels, but it can also point to the difficulty and subjectivity of the task (Plank, 2022), unveiling ambiguous, unclear or underspecified word usages. A tendency for a word to have lower levels of agreement can indicate a vaguer meaning (McCarthy et al., 2016) or other word characteristics that hinder its comprehension.

Low *intra*-annotator (Abercrombie et al., 2023)

agreement is also worth exploring. Schober (2005) found evidence of "conceptual misalignment" (i.e., speakers having different mental representations of words) in surveys: Respondents taking the same survey twice were more likely to change their answers when provided with clarifications such as word definitions in the second iteration. This shows that there had been a misunderstanding the first time around. However, this kind of data is expensive to obtain and it requires formulating hypotheses about the words that are likely to be misunderstood.

4 Methods

We present here metrics, tools, features and methods that can be useful for PPWU prediction, for example because they have been shown to correlate with, or help predict, factors linked to PPWUs. We do not discuss supervised models tailored to the specific NLP tasks.

4.1 Language Model-derived measures

Language models (LMs) are trained with large amounts of text and can provide a measure of the likelihood of sentences or of words in context. Different LM-derived measures can be used to estimate the predictability of a word.

One commonly used measure is word **surprisal**, calculated as the negative log probability of a word occurrence. Surprisal theory (Hale, 2001) states that the cognitive effort required to process a word is proportional to its surprisal. This relationship has been largely studied, and surprisal has been found to be a generally good predictor of words' reading times (RT) (Smith and Levy, 2013; Goodkind and Bicknell, 2018); although it seems to better reflect RTs of lower verbal intelligence profiles (Haller et al., 2024). The choice of LM, however, is an important one. Surprisal values from large LMs with lower perplexities provide worse RT estimations than smaller models (Oh and Schuler, 2023).[9]

Another LM-derived measure is **entropy**. While surprisal is a measure of how unexpected a given word is in its context, entropy is calculated on the probability distribution over the vocabulary, elicited by the preceding context of a specific word position. It can be interpreted as the difficulty of predicting the continuation of a context. Pimentel et al. (2023) use entropy as an operationalization of anticipation and find that it can be a better predictor of RTs than surprisal.

One limitation of these measures for PPWU detection is that they are not only sensitive to (some) semantic anomalies but also to misspellings and syntactic deviations. One promising first step toward solving this problem, so far at the text level, is the contrastive perplexity score (Todd et al., 2020). The difference in perplexity between two LMs (one of which has been fine-tuned on a specific domain) gives information about the nature of the anomaly found (i.e., semantic vs non-semantic).

4.2 Word vector representations

Vector representations of words have been used for a long time in NLP. Static representations from Vector Space Models or word2vec (Mikolov et al., 2013), where a word is assigned a unique vector, have been shown to reflect multiple aspects of words' semantics. For instance, they can be used for identifying synchronic and diachronic lexical variation (Hamilton et al., 2016; Schlechtweg et al., 2019; Gonen et al., 2020), idioms (Peng and Feldman, 2017), hypernyms (Santus et al., 2014) and, in their multilingual form, false friends (Palmero Aprosio et al., 2020) (Section 2.1).

Contextualized word embeddings from Transformer-based language models like BERT (Devlin et al., 2019) additionally allow to represent word semantics at the token level and obtain good results on context-sensitive tasks (Section 2.2) such as novel metaphor detection (Pedinotti et al., 2021) and WSD (Wiedemann et al., 2019). The similarity between a target word and its context has been used to predict eye-tracking features (Salicchi et al., 2023). At the same time, these representations have been shown to encode rich out-of-context lexico-semantic information, such as a word's polysemy level, intensity, complexity and figurativeness (Garí Soler and Apidianaki, 2020; Xypolopoulos et al., 2021; Lyu et al., 2023).

4.3 Cognitive and neurolinguistic measures

Although costly to obtain, psycholinguistic and neurolinguisic data at the word level, as measured through **eye-tracking** and electroencefalography (EEG), can also help identify PPWUs. See Table 2 for datasets annotated with related metrics.

Longer eye fixations and reading times are typically interpreted as an indication of higher cognitive load (Just and Carpenter, 1980; Kintsch et al., 1975). Some eye tracking-derived measures, such

[9] The authors hypothesize that this is due to the vastly larger amount of training data compared to what humans are exposed to.

as gaze duration and first fixation duration, are affected by word frequency and predictability (Inhoff, 1984). Ambiguity also plays a role: Eye fixation times are longer on ambiguous words which have equally likely meanings (Rayner and Duffy, 1986) or for which previous context instantiated a less frequent meaning (Sereno et al., 1992).

The **N400** is an event-related potential signal in the brain consisting of a negative peak occurring about 400 milliseconds after a stimulus. It has been observed as a reaction to all sorts of semantically surprising input (Kutas and Hillyard, 1980; Kutas and Federmeier, 2011). Marked N400 waves have also been found in the processing of metaphorical language (Coulson, 2008). N400 has been shown to reflect word predictability modeled as LM surprisal (Michaelov et al., 2024), and evidence from joke processing shows that it is also sensitive to semantic plausibility operationalized as the similarity of a word with its preceding context (Xu et al., 2024). However, thematic role violations (e.g., between a verb and its argument: "Every morning at breakfast the eggs would eat...") seem to ellicit instead a P600 effect, typically linked with syntactic violations and ambiguities, but no N400 effect (Kuperberg, 2007).

4.4 Linguistic features

The earliest approaches to solving the NLP tasks presented in Section 2 often relied on different kinds and combinations of carefully selected linguistic features. For example, frequency measures (raw, comparative, or diachronic) can be useful for neologism detection (Garcia-Fernandez et al., 2011) and terminology extraction (Rigouts Terryn et al., 2020). PoS and syntactic features have been used to predict disagreement in sense annotations (Martínez Alonso et al., 2015). In lexical complexity prediction, multiple statistical, formal and psycholinguistic features (e.g., word length and concreteness) are also often used (Desai et al., 2021). Ngram information can be employed for neologism (Falk et al., 2014) and code-switching detection (Kevers, 2022), and word co-occurrence information for semantic content quantification (Herbelot and Ganesalingam, 2013). Finally, information on words' selectional preferences has been helpful for metaphor detection (Haagsma and Bjerva, 2016).

4.5 Other approaches

While it is not possible to present all existing methods here, there are other, less widespread approaches that are worth mentioning. For example, the **uncertainty of automatic sense annotations** from WSD models (Liu and Liu, 2023) can be used to find ambiguous word usages as well as to identify words with a tendency to present disambiguation difficulties. **Anomaly** or **outlier detection** methods, although they are most often applied to texts and not at the word level (Ruff et al., 2019; Arora et al., 2021), have been used to detect novel as well as figurative and metaphoric word usages (Sasaki and Shinnou, 2012; Bejan et al., 2023) and idioms (Feldman and Peng, 2013). **Topic modeling** has been used for novel word sense detection (Lau et al., 2014) and idiom detection (Peng et al., 2014); and **sentiment analysis** can be useful for controversial topic detection (Choi et al., 2010). There has also been work evaluating the ability of generative Large LMs to detect ambiguities, not restricted to lexical ones (Liu et al., 2023); but LLMs' effectiveness to detect different kinds of PPWUs remains unexplored.

5 Open Directions and Challenges

As we have shown, there is a substantial amount of work on detecting individual phenomena or word characteristics associated with PPWUs. For the study of communicative success and failure, as well as for the outlined applications of PPWU detection, we argue that it is worth aspiring to address PPWUs in their full diversity, encompassing the multiple factors presented, but no such approach currently exists.[10] The present survey aims to provide an initial overarching perspective that sets the ground for future work.

One key reason for the lack of a comprehensive approach is the **complexity and the varied nature** of PPWUs. The most effective solution may involve a combination of multiple approaches tailored to specific purposes, audiences, situations, or PPWU types. A big gap preventing progress is the **scarcity of large, good-quality datasets** annotated with real PWUs. Annotation of WMNs is helpful but costly; recent efforts striving toward their semi-automatic annotation (Garí Soler et al., 2025b) will contribute to a better understanding of the characteristics of PWUs and their rate of

[10] Lexical complexity prediction is closely related, but it has a narrower scope. It typically focuses on concept-related difficulties that can cause non-understanding. Our definition of PPWUs is broader and more context-dependent, including usages that cause confusion and disagreement, and considers phenomena such as humor, ambiguity, and lexical errors.

occurrence, and will provide more data that could eventually serve to train models for (P)PWU detection. Existing metrics and methodology (Section 4) typically address a particular phenomenon, but it is not always clear what kinds of PPWUs they may detect, which ones they may fail to capture, and what kind of false positives they may propose. For example, LM-derived measures can find unexpected usages of words, which is practical for, e.g., lexical error detection, but would probably not be useful to identify contextual underspecification problems.

Another difficulty lies in the **subjectivity of PPWUs**. As mentioned in Section 1, what constitutes an actual PWU depends on various characteristics of both the speaker and the listener, such as their age, language proficiency, and any potential language impairments or pathologies. It is therefore important to take annotator characteristics into account when collecting data for PPWU detection, and more data and studies are needed to determine the types of usages that are problematic for each type of audience. Differences are not necessarily restricted to specific groups – individual variation can occur even within a community, such as among language learners (Degraeuwe and Goethals, 2024). While the individual component should not be neglected, and it may be desirable to restrict the scope to usages problematic for a specific population, it is also possible to identify usages that are problematic across communities. Such usages may involve errors, underspecification, or jokes; where the issue comes from the production side and depends less on the listener's characteristics.

Unsignaled PPWUs present an additional challenge. Not all PPWUs result in communication breakdowns: Many misunderstandings and disagreements are not signaled, either because they go undetected or because they do not pose a real problem to continue communication. A consequence of this is that when collecting evidence of PPWUs in conversation, an undetermined number of PPWUs are prone to be overlooked. Identifying **keywords** within a dialog may help determine the words that are critical for the conversation to move forward. In a conversational system, this could help restrict the words for which clarification is needed. It is possible that PPWUs worthy of being signaled are more likely to be located in the parts of utterances that introduce new information (i.e., the comment, rather than the topic or theme, in information structure (Lambrecht, 1994)).

Although the perspective proposed here is intended to be language-independent, there may also be **language- and culture-specific** causes of PPWUs, and ways of adapting metrics and tools to specific languages. Whether a PPWU is signaled in a conversation may have a cultural component and be influenced by different politeness norms, potentially causing differences in the frequency with which they are observed across languages.

6 Conclusion

We have introduced the notion of Potentially Problematic Word Usages (PPWUs) and reviewed work from various disciplines to provide a comprehensive perspective of related linguistic factors, methods for their detection, and the kinds of data that can be used for their study. We have also discussed challenges to take into account when working with PPWUs, and identified areas for future work.

PPWUs are a complex and understudied phenomenon. They are affected by multiple factors, both linguistic (word and contextual properties) and non-linguistic (e.g., interlocutor characteristics and differences between them). More work is needed to broaden our understanding of PPWUs, their underlying causes and their prevalence, but datasets containing evidence of miscommunication due to specific word usages are scarce.

One of our long-term goals is to develop methods aimed at capturing PPWUs in their many forms, rather than focusing on just one subtype. With this work, we have laid out a framework organizing existing knowledge and methods, pointing out the diversity of problems that must be considered when tackling PPWUs, in order to serve as a stepping stone for future research on this multifaceted phenomenon.

Limitations

While we have made an effort to provide a comprehensive and cross-disciplinary overview of phenomena, datasets and approaches relevant to the study of problematic word usages, some relevant works may have been missed. Even regardless of the methodology for bibliography search, however, this survey cannot be fully exhaustive. For instance, we have identified and presented discrete factors associated with PPWUs, but word usages can be problematic due to other idiosyncratic reasons that may be hard to classify and detect (see last row of Table 1 for an example). Moreover, it remains

unclear how often each of the presented factors actually leads to a PWU.

In this paper, partly due to space constraints, we have chosen to give only a high-level overview of the different research directions that need to be explored for studying PPWUs. Our primary goal was to establish a conceptual and empirical foundation for the study of PPWUs, which required an extensive synthesis of prior work. While we point to several avenues for future research, taking the first concrete steps, such as carrying out corpus-based studies that could propose a complete typology of PPWUs, is left for future work.

Finally, we acknowledge that, if taken naively, the definition of PPWUs is very broad and many words in a text could be considered to be "potentially problematic." However, this is precisely why we argue that PPWU identification must be audience- and context-sensitive. Bringing these factors together under a single perspective is crucial, because in real-life interaction, problems may arise from different sources, and effectively determining the underlying cause requires considering them jointly.

Acknowledgments

We thank all anonymous reviewers for their feedback. This research was partially funded by the Agence Nationale de la Recherche SINNet project (ANR-23-CE23-0033-01). Additional support was provided by the ANR under the France 2030 program PRAIRIE (ANR-23-IACL-0008).

References

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.

B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part II 12*, pages 277–288. Springer.

Maha H Alhaysony. 2017. Strategies and Difficulties of Understanding English idioms: A Case Study of Saudi University EFL Students. *International Journal of English Linguistics*, 7(3):70–84.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Eef Ameel, Gert Storms, Barbara C Malt, and Steven A Sloman. 2005. How bilinguals solve the naming problem. *Journal of memory and language*, 53(1):60–80.

Mohd Zeeshan Ansari, MM Sufyan Beg, Tanvir Ahmad, Mohd Jazib Khan, and Ghazali Wasim. 2021. Language Identification of Hindi-English tweets using code-mixed BERT. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 248–252. IEEE.

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A Resource and Analyses on Edits in Instructional Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Talita Anthonio and Michael Roth. 2020. What Can We Learn from Noun Substitutions in Revision Histories? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1359–1370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Udit Arora, William Huang, and He He. 2021. Types of Out-of-Distribution Texts and How to Detect Them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ekaterina Artemova and Barbara Plank. 2023. Low-resource Bilingual Dialect Lexicon Induction with Large Language Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 371–385, Tórshavn, Faroe Islands. University of Tartu Library.

Ghislain Atemezing, Benjamin Icard, and Paul Egré. 2022. Vague Terms in SKOS to detect vagueness in textual documents (Version v2).

Aqil M Azmi, Manal N Almutery, and Hatim A Aboalsamh. 2019. Real-word errors in Arabic texts: A better algorithm for detection and correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1308–1320.

Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. 2023. You Told Me That Joke Twice: A Systematic Investigation of Transferability and Robustness of Humor Detection Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13701–13715, Singapore. Association for Computational Linguistics.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 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. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita @ EVALITA2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Matei Bejan, Andrei Manolache, and Marius Popescu. 2023. AD-NLP: A Benchmark for Anomaly Detection in Natural Language Processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10766–10778, Singapore. Association for Computational Linguistics.

Gemma Boleda, Marco Baroni, The Nghia Pham, and Louise McNally. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 35–46, Potsdam, Germany. Association for Computational Linguistics.

Sophie A Booton, Alex Hodgkiss, Sandra Mathers, and Victoria A Murphy. 2022. Measuring knowledge of multiple word meanings in children with english as a first and an additional language and the relationship to reading comprehension. *Journal of Child Language*, 49(1):164–196.

Daniel Bravo-Candel, Jésica López-Hernández, José Antonio García-Díaz, Fernando Molina-Molina, and Francisco García-Sánchez. 2021. Automatic Correction of Real-Word Errors in Spanish Clinical Texts. *Sensors*, 21(9):2893.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022a. DiBiMT: A Novel Benchmark for Measuring Word Sense Disambiguation Biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Niccolò Campolungo, Tommaso Pasini, Denis Emelin, and Roberto Navigli. 2022b. Reducing Disambiguation Biases in NMT by Leveraging Explicit Word Sense Information. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4824–4838, Seattle, United States. Association for Computational Linguistics.

Jonathan Charteris-Black. 1998. Compound Nouns and the Acquisition of English Neologisms. *ERIC Document Reproduction service No. ED427535*.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Yejin Cho, Juan Diego Rodriguez, Yifan Gao, and Katrin Erk. 2020. Leveraging WordNet Paths for Neural Hypernym Prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3007–3018, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yoonjung Choi, Yuchul Jung, and Sung-Hyon Myaeng. 2010. Identifying Controversial Issues and their Subtopics in News Articles. In *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2010, Hyderabad, India, June 21, 2010. Proceedings*, pages 140–153. Springer.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. Citeseer.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel Word-sense Identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Seana Coulson. 2008. Metaphor comprehension and the brain. *The Cambridge handbook of metaphor and thought*, pages 177–194.

Erin E Crockett, Monique MH Pollmann, and Ana P Olvera. 2022. You just don't get it: The impact of misunderstanding on psychological and physiological health. *Journal of Social and Personal Relationships*, 39(9):2847–2868.

D Alan Cruse. 1977. The pragmatics of lexical specificity. *Journal of linguistics*, 13(2):153–164.

Andrew V Danilov, Zilya Nuretdinova, Zulfat Miftakhutdinov, Elvira Sharifullina, Nazym Kydyrbayeva, and Tat'Yana Soldatkina. 2023. Polysemy as a Complexity Predictor in school textbooks. In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, pages 721–725. IEEE.

Jasper Degraeuwe and Patrick Goethals. 2024. LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10432–10447, Torino, Italia. ELRA and ICCL.

Abhinandan Tejalkumar Desai, Kai North, Marcos Zampieri, and Christopher Homan. 2021. LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 548–553, 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.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out Conventionalized Metaphors: A Corpus of Novel Metaphor Annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Pedro J Chamizo Domínguez and Brigitte Nerlich. 2002. False friends: their origin and semantics in some selected languages. *Journal of pragmatics*, 34(12):1833–1849.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 128–135, New York City, USA. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on Word Senses and Word Usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18, Suntec, Singapore. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring Word Meaning in Context. *Computational Linguistics*, 39(3):511–554.

Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4337–4344, Reykjavik, Iceland. European Language Resources Association (ELRA).

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *International conference on intelligent text processing and computational linguistics*, pages 435–446. Springer.

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

SzeWang Fong, Dmitri Roussinov, and David B Skillicorn. 2008. Detecting Word Substitutions in Text. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1067–1076.

Peter Freebody and Richard C Anderson. 1983. Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, 15(3):19–39.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it Written? Automatically Determining Publication Dates. In *String Processing and Information Retrieval: 18th International Symposium, SPIRE 2011, Pisa, Italy, October 17-21, 2011. Proceedings 18*, pages 221–236. Springer.

Aina Garí, Marianna Apidianaki, and Alexandre Allauzen. 2018. A comparative study of word embeddings and other features for lexical complexity detection in French. In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 499–508, Rennes, France. ATALA.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT Knows Punta Cana is not just beautiful, it's gorgeous: Ranking Scalar Adjectives with Contextualised Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2021a. Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Aina Garí Soler and Marianna Apidianaki. 2021b. Scalar Adjective Identification and Multilingual Ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2022. One Word, Two Sides: Traces of Stance in Contextualized Word Representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3950–3959, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring Lexico-Semantic Alignment in Debates with Contextualized Word Representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 50–63, Toronto, Canada. Association for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2025b. Toward the Automatic Detection of Word Meaning Negotiation Indicators in Conversation. *Accepted at Findings of the Association for Computational Linguistics: EMNLP 2025*.

Aina Garí Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. 2025a. The NeWMe Corpus: A gold standard corpus for the study of Word Meaning Negotiation. *PREPRINT (Version 1) available at Research Square*.

Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. Identifying Cross-Cultural Differences in Word Usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, 1(1):1–27.

Edward Gibson, Caitlin Tan, Richard Futrell, Kyle Mahowald, Lars Konieczny, Barbara Hemforth, and Evelina Fedorenko. 2017. Don't Underestimate the Benefits of Being Misunderstood. *Psychological science*, 28(6):703–712.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Stefan Th Gries, Marlies Jansegers, and Viola G Miglio. 2020. Quantitative methods for corpus-based contrastive linguistics. In *New approaches to contrastive linguistics: empirical and mathodological challenges*, volume 336, pages 53–84. de Gruyter.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting novel metaphor using selectional preference information. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 10–17, San Diego, California. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Patrick Haller, Lena Bolliger, and Lena Jäger. 2024. Language models emulate certain cognitive profiles: An investigation of how predictability measures interact with individual differences. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7878–7892, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Viktor Hangya, Qianchu Liu, Dario Stojanovski, Alexander Fraser, and Anna Korhonen. 2021. Improving Machine Translation of Rare and Unseen Word Senses. In *Proceedings of the Sixth Conference on Machine Translation*, pages 614–624, Online. Association for Computational Linguistics.

Anna Hätty, Dominik Schlechtweg, Michael Dorna, and Sabine Schulte im Walde. 2020. Predicting Degrees of Technicality in Automatic Terminology Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2883–2889, Online. Association for Computational Linguistics.

Saengchan Hemchua, Norbert Schmitt, et al. 2006. An analysis of lexical errors in the English compositions of Thai learners. *PROSPECT-ADELAIDE-*, 21(3):3.

Aurélie Herbelot and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, Sofia, Bulgaria. Association for Computational Linguistics.

Graeme Hirst, David St-Onge, et al. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463, Baltimore, Maryland. Association for Computational Linguistics.

Ana Werkmann Horvat, Marianna Bolognesi, Jeannette Littlemore, and John Barnden. 2022. Comprehension of different types of novel metaphors in monolinguals and multilinguals. *Language and Cognition*, 14(3):401–436.

Benjamin Icard, Ghislain Atemezing, and Paul Égré. 2022. VAGO: un outil en ligne de mesure du vague et de la subjectivité. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (PFIA 2022)*, pages 68–71.

Albrecht Werner Inhoff. 1984. Two stages of word processing during eye fixations in the reading of prose. *Journal of verbal learning and verbal behavior*, 23(5):612–624.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic Identification of Cognates and False Friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

Suzanne Irujo. 1986. Don't put your Leg in your Mouth: Transfer in the Acquisition of Idioms in a Second Language. *tesol Quarterly*, 20(2):287–304.

Ann Irvine and Chris Callison-Burch. 2017. A Comprehensive Analysis of Bilingual Lexicon Induction. *Computational Linguistics*, 43(2):273–310.

Maarten Janssen. 2012. NeoTag: a POS Tagger for Grammatical Neologism Detection. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2118–2124, Istanbul, Turkey. European Language Resources Association (ELRA).

Scott Jarvis. 2011. Conceptual transfer: Crosslinguistic effects in categorization and construal. *Bilingualism: Language and cognition*, 14(1):1–8.

Heng Ji and Kevin Knight. 2018. Creative Language Encoding under Censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 23–33, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3006–3012, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Diptesh Kanojia, Prashant Sharma, Sayali Ghodekar, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2021. Cognition-aware Cognate Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3281–3292, Online. Association for Computational Linguistics.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

J Kenneth Logan and Michael J Kieffer. 2017. Evaluating the role of polysemous word knowledge in reading comprehension among bilingual adolescents. *Reading and Writing*, 30:1687–1704.

Laurent Kevers. 2022. CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121, Marseille, France. European Language Resources Association.

Boaz Keysar and Anne S Henly. 2002. Speakers' overestimation of their effectiveness. *Psychological Science*, 13(3):207–212.

Walter Kintsch, Ely Kozminsky, William J Streby, Gail McKoon, and Janice M Keenan. 1975. Comprehension and recall of text as a function of content variables. *Journal of verbal learning and verbal behavior*, 14(2):196–214.

Annette Klosa and Harald Lüngen. 2018. New German words: Detection and description. In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts*, pages 559–569, Ljubljana. Znanstvena založba Filozofske fakultete Univerze v Ljubljani/Ljubljana University Press, Faculty of Arts.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-Informed Specificity and Semantic Plausibility for Dialogue Generation. 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 3456–3466, Minneapolis, Minnesota. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 Task 5: Evaluating Phrasal Semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for Clarification Question Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Gina R Kuperberg. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research*, 1146:23–49.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62(1):621–647.

Marta Kutas and Steven A Hillyard. 1980. Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological psychology*, 11(2):99–116.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic Semantic Change Dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Vicky Tzuyin Lai, Tim Curran, and Lise Menn. 2009. Comprehending conventional and novel metaphors: An ERP study. *Brain research*, 1284:145–155.

George Lakoff. 1993. *The contemporary theory of metaphor*, page 202–251. Cambridge University Press.

Knud Lambrecht. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge university press.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270, Baltimore, Maryland. Association for Computational Linguistics.

John Lee and Chak Yan Yeung. 2018. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Els Lefever, Sofie Labat, and Pranaydeep Singh. 2020. Identifying Cognates in English-Dutch and French-Dutch by means of Orthographic Information and Cross-lingual Word Embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4096–4101, Marseille, France. European Language Resources Association.

Adrienne Lehrer. 2003. Understanding trendy neologisms. *Italian Journal of Linguistics*, 15:369–382.

Shiri Lev-Ari. 2015. Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in psychology*, 5:1546.

Linlin Li and Caroline Sporleder. 2010. Linguistic Cues for Distinguishing Literal and Non-Literal Usages. In *Coling 2010: Posters*, pages 683–691, Beijing, China. Coling 2010 Organizing Committee.

Maya R Libben and Debra A Titone. 2008. The multidetermined nature of idiom processing. *Memory & cognition*, 36:1103–1121.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Zhu Liu and Ying Liu. 2023. Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3963–3977, Toronto, Canada. Association for Computational Linguistics.

Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Steven G Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50:826–833.

Qing Lyu, Marianna Apidianaki, and Chris Callison-burch. 2023. Representation of Lexical Stylistic Features in Language Models' Embedding Space. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 370–387, Toronto, Canada. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Barbara C Malt, Steven A Sloman, and Silvia P Gennari. 2003. Universality and language specificity in object naming. *Journal of memory and language*, 49(1):20–42.

Carolyn B Marks, Marleen J Doctorow, and Merlin C Wittrock. 1974. Word Frequency and Reading Comprehension. *The Journal of Educational Research*, 67(6):259–262.

Ron Martinez and Victoria A Murphy. 2011. Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *Tesol Quarterly*, 45(2):267–290.

Héctor Martínez Alonso, Anders Johannsen, Oier Lopez de Lacalle, and Eneko Agirre. 2015. Predicting word sense annotation agreement. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 89–94, Lisbon, Portugal. Association for Computational Linguistics.

Marina Mattheoudakis and Paschalia Patsala. 2007. English-Greek false friends: Now they are, now they aren't. In *8th International Conference of Greek Linguistics. University of Ioannina, August 30th-September 7th*.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Computational Linguistics*, 42(2):245–275.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Automatic Identification of Infrequent Word Senses. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1220–1226, Geneva, Switzerland. COLING.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2024. Strong Prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of language*, 5(1):107–135.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint:1301.3781v3*.

Ruslan Mitkov, Viktor Pekar, Dimitar Blagoev, and Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine translation*, 21:29–53.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the Second Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Syrielle Montariol, Aina Garí Soler, and Alexandre Allauzen. 2019. Exploring sentence informativeness. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 303–312, Toulouse, France. ATALA.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and Interpretable Semantic Change Detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Jenny Myrendal. 2015. *Word meaning negotiation in online discussion forum communication*. Ph.D. thesis, University of Gothenburg.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse studies*, 21(3):317–339.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Daiki Nishihara and Tomoyuki Kajiwara. 2020. Word Complexity Estimation for Japanese Lexical Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3114–3120, Marseille, France. European Language Resources Association.

Bill Noble, Staffan Larsson, and Jenny Myrendal. 2025. Misunderstanding the Concrete, Disagreeing About the Abstract: A Closer Look at Word Meaning Negotiation Triggers. In *Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue – Full Papers*, pages 81–91, Bielefeld, Germany. SEMDIAL.

Bill Noble, Kate Viloria, Staffan Larsson, and Asad Sayeed. 2021. What do you mean by negotiation? Annotating social media discussions about word meaning. In *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021)*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Nadia Nouri and Badia Zerhouni. 2018. Lexical Frequency Effect on Reading Comprehension and Recall. *Arab World English Journal (AWEJ) Volume*, 9.

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Jenny A Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Adaptive Complex Word Identification through False Friend Detection. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 192–200.

Jinkyoung Park and Yuah V Chon. 2019. EFL Learners' Knowledge of High-frequency Words in the Comprehension of Idioms: A Boost or a Burden? *RELC journal*, 50(2):219–234.

Aneta Pavlenko. 2009. *The bilingual mental lexicon: Interdisciplinary approaches*, volume 70. Multilingual Matters.

Aneta Pavlenko and Barbara C Malt. 2011. Kitchen Russian: Cross-linguistic differences and first-language object naming by Russian–English bilinguals. *Bilingualism: Language and Cognition*, 14(1):19–45.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(Feb):79–92.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? Testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Peng and Anna Feldman. 2017. Automatic idiom recognition with word embeddings. In *Information Management and Big Data: Second Annual International Symposium, SIMBig 2015, Cusco, Peru, September 2-4, 2015, and Third Annual International Symposium, SIMBig 2016, Cusco, Peru, September 1-3, 2016, Revised Selected Papers 2*, pages 17–29. Springer.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2014. Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2019–2027, Doha, Qatar. Association for Computational Linguistics.

Sandro Pezzelle and Raquel Fernández. 2023. Semantic Adaptation to the Interpretation of Gradable Adjectives via Active Linguistic Interaction. *Cognitive Science*, 47(2):e13248.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

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

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger P. Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times. *Transactions of the Association for Computational Linguistics*, 11:1624–1642.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the Means for Clarification in Dialogue. In R. Smith and J. van Kuppevelt, editors, *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Technology*, pages 235–255. Kluwer Academic Publishers.

Marlou Rasenberg, Wim Pouw, Asli Özyürek, and Mark Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, 12(1):19111.

Daniel Raušer. 2017. Selected English-Czech False Friends and Their Use in the Works of Some Czech Students. *Caracteres*, Vol. 6.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Sebastian Reimann and Tatjana Scheffler. 2024. When is a Metaphor Actually Novel? Annotating Metaphor Novelty in the Context of Automatic Metaphor Detection. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada. Association for Computational Linguistics.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Roth, Talita Anthonio, and Anna Sauer. 2022. SemEval-2022 Task 7: Identifying Plausible Clarifications of Implicit and Underspecified Phrases in Instructional Texts. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1039–1049, Seattle, United States. Association for Computational Linguistics.

Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.

Lavinia Salicchi, Emmanuele Chersoni, and Alessandro Lenci. 2023. A study on surprisal and semantic relatedness for eye-tracking data prediction. *Frontiers in Psychology*, 14:1112365.

Pratip Samanta and Bidyut B. Chaudhuri. 2013. A simple real-word error detection and correction using local word bigram and trigram. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 211–220, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual Code-switching Identification via LSTM Recurrent Neural Networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas. Association for Computational Linguistics.

Enrico Santus, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing Hypernyms in Vector Spaces with Entropy. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Minoru Sasaki and Hiroyuki Shinnou. 2012. Detection of Peculiar Word Sense by Distance Metric Learning with Labeled Examples. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 601–604, Istanbul, Turkey. European Language Resources Association (ELRA).

Wafa Ismail Saud. 2018. Lexical Errors of Third Year Undergraduate Students. *English Language Teaching*, 11(11):161–168.

Antoinette Schapper and Maria Koptjevskaja-Tamm. 2022. Introduction to special issue on areal typology of lexico-semantics. *Linguistic Typology*, 26(2):199–209.

Yves Scherrer. 2007. Adaptive String Distance Measures for Bilingual Dialect Lexicon Induction. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 55–60, Prague, Czech Republic. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2019. Attentive Mimicking: Better Word Embeddings by Attending to Informative Contexts. 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 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi Shared Task: Median Judgment Classification & Mean Disagreement Ranking with Ordinal Word-in-Context Judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael F Schober. 2005. Conceptual Alignment in Conversation. *Other minds: How humans bridge the divide between self and others*, pages 239–252.

Gregory Schraw, Woodrow Trathen, Ralph E Reynolds, and Richard T Lapan. 1988. Preferences for Idioms: Restrictions due to Lexicalization and Familiarity. *Journal of Psycholinguistic Research*, 17:413–424.

Marc Schulder and Eduard Hovy. 2014. Metaphor Detection through Term Relevance. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 18–26, Baltimore, MD. Association for Computational Linguistics.

Sara C Sereno, Jeremy M Pacht, and Keith Rayner. 1992. The effect of meaning frequency on processing lexically ambiguous words: Evidence from eye fixations. *Psychological Science*, 3(5):296–301.

Nadia A Shalaby, Noorchaya Yahya, and Mohamed El-Komi. 2009. Analysis of Lexical Errors in Saudi College Students' Compositions. *Journal of the Saudi Association of Languages and Translation*, 2(3):65–93.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving Hypernymy Detection with an Integrated Path-based and Distributional Method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. NewsEdits: A News Article Revision Dataset and a Novel Document-Level Reasoning Challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in Context: The IDIX Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. Amsterdam: John Benjamins.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Graham Todd, Catalin Voss, and Jenny Hong. 2020. Unsupervised Anomaly Detection in Parole Hearings using Language Models. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 66–71, Online. Association for Computational Linguistics.

Kathryn K Toffolo, Edward G Freedman, and John J Foxe. 2022. Evoking the N400 Event-related Potential (ERP) Component Using a Publicly Available Novel Set of Sentences with Semantically Incongruent or Congruent Eggplants (Endings). *Neuroscience*, 501:143–158.

Robert Van Rooij. 2011. Vagueness and Linguistics. In *Vagueness: A guide*, pages 123–170. Springer.

Evangeline Marlos Varonis and Susan Gass. 1985. Non-native/non-native conversations: A model for negotiation of meaning. *Applied linguistics*, 6(1):71–90.

Leo Wanner, Serge Verlinde, and Margarita Alonso Ramos. 2013. Writing assistants and automatic lexical error correction: word combinatorics. *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of eLex 2013*, pages 472–487.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 161–170, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Rodrigo Wilkens, Alessandro Dalla Vecchia, Marcely Zanon Boito, Muntsa Padró, and Aline Villavicencio. 2014. Size does not matter. Frequency does. A study of features for measuring lexical complexity. In *Advances in Artificial Intelligence–IBERAMIA 2014: 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24-27, 2014, Proceedings 14*, pages 129–140. Springer.

Haoyin Xu, Masaki Nakanishi, and Seana Coulson. 2024. Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking Clarification Questions in Knowledge-Based Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

Christos Xypolopoulos, Antoine Tixier, and Michalis Vazirgiannis. 2021. Unsupervised Word Polysemy Quantification with Multiresolution Grids of Contextual Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401, Online. Association for Computational Linguistics.

Yusuf Yaylaci and Arman Argynbayev. 2014. English-Russian False Friends in ELT Classes with Intercultural Communicative Perspectives. *Procedia-Social and Behavioral Sciences*, 122:58–64.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The Global Anchor Method for Quantifying Linguistic Shifts and Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Appendix

B PPWU contributing factors: Diagram

Cause	Example	Source
Low frequency	*Yes, I am a **byword** to them.*	Shardlow et al. (2020)
Neologism	*– I, yeah, I-m, yeah, I do aerobics, uh, **step classes** and, uh* *– Step classes?* *– toning classes, yes.* (At the time of speaking, 1992, stepping classes were a new concept)	Godfrey et al. (1992); Garí Soler et al. (2025a)
Lexical variation	*chips* denotes different things in UK and American English, and *lorry* is mainly exclusively used in the UK.	
Complex meaning	*It is unlikely that morphological changes during development led to impairment of spatial learning and motor coordination, and morphological alterations in the cytoarchitecture of the **hippocampus** and cerebellum were not observed* (...)	Shardlow et al. (2020)
Vagueness	*– Do you genuinely believe that the less palatable parts of TRP* [The Red Pill] *are less **effective** than the positive parts?* *– You'll have to define "effective" for me and give me a specific example or two* (...)*. Many forms of abuse, manipulation, and deception might be described as "effective", depending on what your goal is and whether or not you're a psychopath.*	Tan et al. (2016); Garí Soler et al. (2025a)
Generality	*– I wonder what kind of **care** you get when you're hospitalised for appendicitis, gallstone, ileus, inflammation of the pancreas or inflammation of the intestines.* (...) *– What do you mean by care? I was brought medicine and received help with mixing the nutritional drink which was replacing food during the time my intestines were resting. When I started feeling better and could move around I had to fix that myself.*	Myrendal (2015)
Connotatively loaded term	*–* (...) ***agnosticism** postulates that the existence or nonexistence of god is beyond our knowledge or ability to gather it.* (...) *– Not necessarily, or in most cases. Agnosticism argues that it is not possible at a given moment in time to know absolutely, but then we don't know anything absolutely. Moreover, that that's ok, we'll work with what we have.* (...) *–* (...) *That's kind of a flimsy sort of agnosticism* (...)	Tan et al. (2016); Garí Soler et al. (2025a)
False friend	*We usually go to a **magazine** to buy milk.* (Russian магазин (magazin) or French *magasin* mean *shop, store*)	Yaylaci and Argynbayev (2014)
Partial cognate	English *blank* and its equivalent in several Romance languages (ES: *blanco*, FR: blanc, PT: branco, IT: bianco, CA: blanc). They can both mean *empty* but the Romance versions also mean *white*.	Domínguez and Nerlich (2002)
Cross-linguistic inequivalence	Russian форточка (fortochka), a specific kind of small window for ventilation common in post-Soviet states.	Pavlenko (2009)
Word from another language	*–* (...) *the truth is that I don't know what the **gringo** fandom is like* (...) *–* (...) *Not sure what you mean by gringo* (...) *– Oh JAJFJWJF SORRY- "gringo" is a way that people who speak spanish refer to people who speak english* (...)	Noble et al. (2021)
Ambiguity or contextual underspecification	*Rooms are classically decorated and **warm***	Jurgens (2014)
Ambiguity or contextual underspecification	*–* (...) *I enjoy **working** on my car* (...) (...) *– Oh I thought you meant "working on my car" as in polishing it and keeping it in super condition as a hobby, not as in "occasional repairs".*	Tan et al. (2016); Garí Soler et al. (2025a)
Novel sense	The use of *bet* to mean *yes*	
Metaphor	*Westerns have a **gladiatorial**, timeless quality.*	Do Dinh et al. (2018)
Lexical error	*Choose which **charter** you want to be .* (correction: character)	Anthonio et al. (2020)
Rare sense	*– Anyone who has a link to a **dirty** Win7 download?* *– What do you mean by dirty?* *– What I meant by 'dirty' was an illegal copy of the operating system.*	Myrendal (2015)
Vandalism	*First , make sure your hamster is familiar with your scent and your **poop** .* (original word: "voice")	Anthonio and Roth (2020)
Semantic plausibility / contradiction	*i understand. i am not sure if i can afford a babysitter, i am a **millionaire***	Ko et al. (2019)
Unclassified	*–* (...) *True **waffles** are crisp on the outside and fluffy on the inside* (...) *– So the definition of waffle changes based on how long you cook it? I happen to enjoy my waffles slightly undercooked, does that mean that they are not waffles?*	Tan et al. (2016); Garí Soler et al. (2025a)

Table 1: Examples of PPWUs.

Task/Annotation	Reference	Language	Resource name (if existing)
Lexical Semantic Change	Schlechtweg et al. (2018)	DE	DURel
	Schlechtweg et al. (2020)	EN, DE, SV, LA	SemEval 2020 Task 1
	Basile et al. (2020)	IT	DIACR-Ita
	Rodina and Kutuzov (2020)	RU	RuSemShift
	Kutuzov et al. (2022)	NO	NorDiaChange
	Chen et al. (2023)	ZH	ChiWUG
Complex Word Identification	Horn et al. (2014)	EN	
	Paetzold and Specia (2016)	EN	CWI–2016
	Yimam et al. (2018)	EN, DE, ES	CWI–2018
	Maddela and Xu (2018)	EN	Word Complexity Lexicon
	Lee and Yeung (2018)	EN	
	Ortiz-Zambranoa and Montejo-Ráezb (2020)	ES	ALexS
	Nishihara and Kajiwara (2020)	JA	
	Shardlow et al. (2021)	EN	CompLex
Idiomatic Expressions	Haagsma et al. (2020)	EN	MAGPIE
	Korkontzelos et al. (2013)	EN	SemEval-2013 Task 5
	Cook et al. (2008)	EN	VNC-Tokens
	Sporleder et al. (2010)	EN	IDIX
Term Extraction	Rigouts Terryn et al. (2020)	EN, FR, NL	ACTER
Scalar Adjective Identification	Garí Soler and Apidianaki (2021b)	EN	SCAL-REL
Vague terms	Atemezing et al. (2022)	EN, FR	
False friends and cognates	Palmero Aprosio et al. (2020)	IT-{EN,FR,DE,ES}	
Bilingual Lexicon Induction	Pavlick et al. (2014)	100 languages	
Code-switching	Solorio et al. (2014)	MSA-DA, EN-ES, EN-ZH, EN-NE	
	Molina et al. (2016)	MSA-DA, EN-ES	
	Kevers (2022)	CO-FR	BDLC
	Nayak and Joshi (2022)	HI-EN	L3Cube-HingCorpus
Novel Word Sense Detection	Cook et al. (2014)	EN	BNC-ukWaC & SiBol/Port
Metaphor Detection	Steen et al. (2010)	EN	VUAMC
Novel Metaphor Detection	Do Dinh et al. (2018)	EN	
Self-paced reading times	Smith and Levy (2013)	EN	Brown Corpus
Self-paced reading times	Futrell et al. (2021)	EN	Natural Stories Corpus
Word-level eye-tracking	Kennedy et al. (2003)	EN	Dundee Corpus
Word-level eye-tracking	Luke and Christianson (2018)	EN	Provo Corpus
N400	Toffolo et al. (2022)	EN	

Table 2: Selection of datasets annotated with linguistic information relevant for PPWUs.

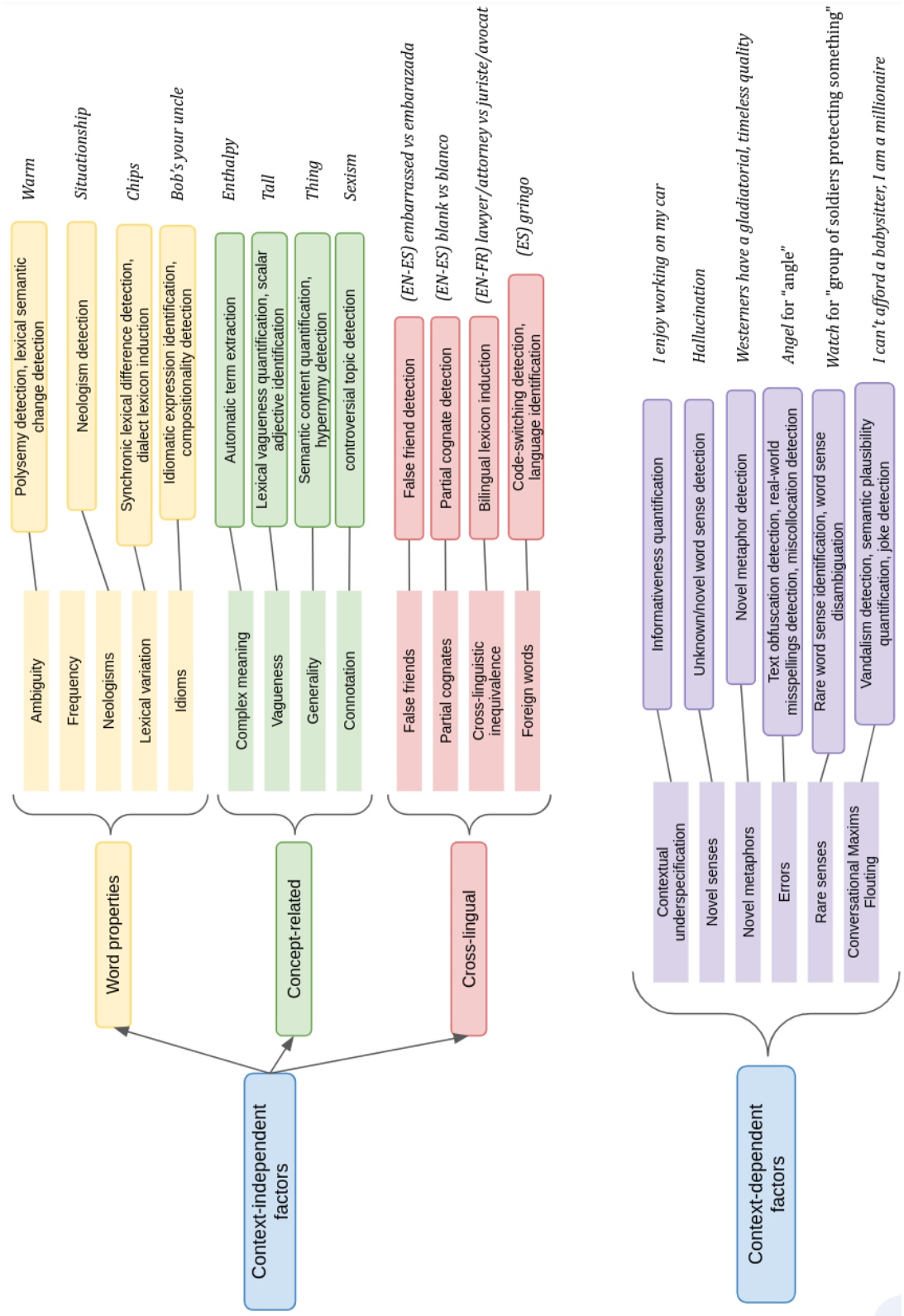

Figure 1: Factors associated with PPWUs and related NLP tasks, with examples.