Unsupervised Methods to Predict Example Difficulty in Word Sense Annotation

Author: Cristina Aceta Moreno

Advisors: Oier López de Lacalle, Eneko Agirre and Izaskun Aldezabal

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua
Language Analysis and Processing

Final Thesis

June 2018

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Hitzen Adiera Desanbiguazioa (HAD) Hizkuntzaren Prozesamenduko (HP) erronkarik handienetakoa da. Frogatu denez, HAD sistema ahalik eta arrakastatsuenak entrenatzeko, oso garrantzitsua da entrenatze-datuetatik adibide (hitzen testuinguru) zailak kentzea, honela emaitzak asko hobetzen baitira. Lan honetan, lehenik, gainbegiratutako ereduak aztertzen ditugu, eta, ondoren, gainbegiratu gabeko bi neurri proposatzen ditugu. Gainbegiratutako ereduetan, adibideen zailtasuna definitzeko, anotatutako corpuseko datuak erabiltzen dira. Proposatzen ditugun bi gainbegiratu gabeko neurrietan, berriz, batetik, aztergai den hitzaren zailtasuna neurtzen da (hitzon Wordnet-eko datuak aztertuta), eta, bestetik, hitzaren agerpenarena (alegia, hitzaren testuinguruarena edo adibidearena). Biak konbinatuta, adibideen zailtasuna ezaugarritzeko eredu bat ere proposatzen da.

Abstract

Word Sense Disambiguation (WSD) is one of the major challenges in Natural Language Processing (NLP). In order to train successful WSD systems, it has been proved that removing difficult examples (words in a context) from the training set improves the performance of these systems. In this work, we first analyze supervised models that, given annotated data, characterize the difficulty of examples. We then propose two unsupervised measures to characterize the difficulty of target words (by analyzing their WordNet data) and occurrences (context sentences), respectively. Combining them, a model able to characterize the difficulty of examples is also presented.

Acknowledgements

First of all, I would like to thank my directors: Dr Oier López de Lacalle, for all the help along these months, for his patience when I struggled with statistics or R and for solving all my doubts every time I had problems (even on holidays and weekends!); Dr Eneko Agirre, for all the contributions and for helping me giving this project a new turn, and Dr Izaskun Aldezabal, for all the support, the ideas and for her encouragement. Thank you to all of you: without all your help, this project would have never become like this.

I also would like to thank Dr Joaquim Llisterri, for being the first professor to ever teach me Language Technologies −by changing his beloved *Fonètica* for *Tecnologies del Llenguatge* that year−, at Universitat Autònoma de Barcelona. Thank you for being so supportive in your classes, and for confirming me that this is the path I want to follow in my life.

Also, I would like to thank my eusko-friends for their unconditional support: Jesús Vera, for trying to cheer me up while I felt overwhelmed; Iñigo Ayestaran, for infecting me with his positivity; Jon Mikel Olmos, for listening to my stories and Shuyuan Cao, for always making me feel better. Of course, I could not forget my Catalan friends: Míriam Chamorro, Jorge Martínez, Brenda Ruiz, Ibra Jabbi. Thank you for not hating me for not being able to visit you as much as I wanted.

Furthermore, thank you very much to my colleagues from SII in my internship at IK4-Tekniker, for cheering me up, listening to me and helping me. I would like to especially thank Kerman López de Calle, for his (eternal) patience while helping me with R, and Dr Izaskun Fernández and Dr Aitor Arnaiz, for offering me flexibility in order to attend my meetings.

And last, but not least, I want to thank my two families (the Basque and the Catalan one) for all their support. And, especially, thank you to my boyfriend, Ander González, for his absolutely unconditional support, for cheering me up, for offering his help even when he had no idea, for listening to all my explanations to see if they had sense, for not leaving me in my reclusion weekends and for being my fan number 1.

Thank you to all of you (and to the ones I have not mentioned). Without you, this project would not have been possible.

Contents

List of Figures

List of Tables

1 Introduction

Currently, the development of successful Word Sense Disambiguation (WSD) systems is a major challenge in the field of Natural Language Processing (NLP). According to López de Lacalle and Agirre (2015b), the accuracy of WSD systems goes from **60% to 70%**, for words with a high number of training examples. This is due, according to Hovy et al. (2006) to two factors: a lack of large annotated corpora of quality and the senses stored in the sense inventories, which are claimed to be **too fine-grained** in most cases.

The creation of WordNet (Miller, Beckwith, et al. 1990) has become a turning point in the access to words and their storage in a database, since, unlike a thesaurus, it organizes words in terms of cognitive and semantic criteria, establishing relations between them and creating a **semantic network** with all the words in the English language. Such is the importance of WordNet that other WordNets have been created for other languages, such as French, German or Spanish. It also has been integrated to other Natural Language Processing databases, such as BabelNet[1] or DBpedia[2].

The way words are related between them makes WordNet one of the most used databases in Natural Language Processing, and has proven to be very useful for many applications in the field, such as **word-sense disambiguation** (WSD) or **machine translation** (MT).

However, WordNet, despite the innovation and the robustness of the concept, does not achieve good results when used to train WSD systems. In this sense, it is **important** to be able to **determine the problems** that can cause WordNet to achieve not-so-good results when applied to WSD in order to solve them and improve the performance of these systems. The **detection** of these problems can help **creating a model able to characterize problematic examples**.

The low performance of WSD systems trained with WordNet senses can be due to the fact that WordNet senses are sometimes **difficult to annotate**, as the results from Passonneau and Carpenter (2014)'s annotation task show. For example[3]:

Word to disambiguate: *tell-v*

Senses for *tell-v*:

1. Express in words
2. Let something be known
3. Narrate or give a detailed account of
4. Give instructions to or direct somebody to do something with authority
5. Discern or comprehend

[1] https://babelnet.org/
[2] http://wiki.dbpedia.org/
[3] The examples are extracted from Passonneau and Carpenter (2014)'s annotation task.

6. Inform positively and with certainty and confidence
7. Give evidence
8. Mark as different

Context sentences:

- Even here, the channel perspective **tells** a somewhat different story.
- I **told** him about canalizing functions.

In this example, the annotator is supposed to assign a sense of the word ***tell-v*** to the context sentences it appears in. The first problem that can be found here is that the senses for ***tell-v*** are difficult to differentiate, since share very **similar ideas**, even **similar words**, such as senses 1 and 3 (*express*, *narrate*). Also, the context sentences do not seem to convey all the **necessary information** in order to be able to assign a sense to the context sentence.

In this thesis, the dataset from the annotation task in Passonneau and Carpenter (2014) will be analysed, in order to be able to detect problematic examples in an annotation task, by taking into account the **agreement** between annotators. Agreement measures appear to be valid **supervised** metrics in order to distinguish problematic examples. In this project, **kappa agreement** and **entropy** of the example will be evaluated as the supervised metrics to characterize the difficulty of words and context sentences/examples, respectively.

Although supervised metrics are useful for analysis, WSD systems require methods to assess difficulty which do not have access to annotation data. Thus, the analysis in this thesis will be centered in **unsupervised** analysis, following a **linguistic approach**, to identify difficult words and context sentences and, in the end, **difficult examples**.

Motivated by the observations on the data (as in the previous example and as it will be observed in Section 1.2), unsupervised metrics for words and context sentences will be proposed: **similarity between sense definitions** for **words** and **probability** for **context sentences**. In order to confirm the validity of the proposed metrics, both supervised and unsupervised metrics will be **correlated** and, then, the proposed unsupervised metrics will be combined and fitted in a **linear model** that is expected to be able to **quantify the difficulty of an example**.

1.1 Important terminology

All along this thesis, some recurrent terminology will be used. As this terminology will correspond to key elements in the analysis, it is important to have it in mind from the beginning.

These key elements, which will be the objects of analysis in this project, are the **target word**, the **context sentence** and the **example**:

- **Target word** (TW). It is the word to disambiguate. For example:
  - "You will be **added** to the mailing list" - TW: ***add***
  - "Can you **help** me?" - TW: ***help***

  The target word is placed in a sentence which, in a disambiguation task, has to be assigned a sense from an inventory of senses. In order to model the difficulty of a target word, the information provided in the context will not be taken into account in the analysis, only the information strictly related to the word and its inventory of senses.

- **Context sentence**. The context in which the target word is introduced. The most **important** fact about the **context sentence** is that, in order to model its difficulty, the information of the target word (more specifically, its difficulty) will not be considered in its analysis. Thus, the analysis for this element will be centered in the information provided by the **sentence**. For example:
  - "**You will be added to the mailing list**"
  - "**Can you help me?**"

- **Example**. An example is an element in whose analysis both the difficulty of the target word and the difficulty of the context sentence are taken into account. For example:
  - "**You will be *added* to the mailing list**" - Example for ***add***
  - "**Can you *help* me?**" - Example for ***help***

  Thus, the difficulty of an example would be the **combination** of the difficulty of the target word and the difficulty of the context sentence.

1.2 Factors that make annotation difficult[4]

In a disambiguation task, the annotators are provided with two sources of information: the **target word** (more specifically, its inventory of **senses**) and its **context sentences**, so these are possibly the elements that have a major impact on the difficulty of disambiguation of a word in context, as it has been pointed out previously by López de Lacalle and Agirre (2015b).

More precisely, it seems that difficult **words** have **senses** that are **similar** between them, and difficult **context sentences** do not provide enough **information** in order to be disambiguated easily. These observations will be discussed in the following sections.

[4]All the examples in this sections are extracted from Passonneau and Carpenter (2014) annotation task.

The following subsections will consist of observations regarding the difficulty related to words and context sentences, which will help to define the experiments that will be performed later on.

1.2.1 Observations on the difficulty of words

Regarding the **word**, it is possible to consider whether it is easy or not by taking into account its inventory of senses, as it has been explored in López de Lacalle and Agirre (2015b). According to their work, and basing their hypotheses on the studies from Yarowsky and Florian (2002), the difficulty in words may come from their distribution and number of senses in the inventory.

Taking into account the considerations mentioned previously, on the one hand, a **word** could be considered as ***easy*** if:

- The senses are **easy to differentiate**. That is, if the senses do **not overlap**. In other words, the boundaries between them are clear. They may, for example, relate to different topics.
- The **inventory of senses is small**. The less senses to choose from, the less difficulty.

However, since it has been proved in López de Lacalle and Agirre (2015b) that the number of senses is not a crucial factor to determine the difficulty of a word (although it is related to it), the main consideration here will be that an ***easy*** word **does not have overlapping senses**.

An example of an ***easy*** word would be ***sense***:

1) A general conscious awareness.
EX: "A sense of security", "A sense of happiness", "A sense of danger", "A sense of self".

2) The meaning of a word or expression; the way in which a word or expression or situation can be interpreted.
EX: "The dictionary gave several senses for the word", "In the best sense charity is really a duty", "The signifier is linked to the signified".

3) The faculty through which the external world is apprehended.
EX: "In the dark he had to depend on touch and on his senses of smell and hearing".

4) Sound practical judgement.
EX: "Common sense is not so common", "He hasn't got the sense God gave little green apples", "Fortunately she had the good sense to run away".

5) A natural appreciation or ability.
EX: "A keen musical sense", "A good sense of timing".

In this case, the provided senses apply to different situations that the other senses cannot define. In other words, the senses do **not overlap** between them. Also, the examples provided help to distinguish senses that may not be clearly defined (for example, 1 and 5).

On the other hand, a **word** can be considered as ***difficult*** if:

- The senses are **difficult to differentiate**. That is, if the senses **overlap**. The topics they relate to may be the same or very similar.
- The **inventory of senses is big**. The more senses to choose from, the higher the difficulty.

However, and as in the previous case, the main consideration here will be that a ***difficult*** word **has overlapping senses**.

An example of a ***difficult*** word would be ***know***:

1) Be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about.
EX: "I know that the President lied to the people", "I want to know who is winning the game!", "I know it's time".

2) Know how to do or perform something.
EX: "She knows how to knit", "Does your husband know how to cook?".

3) Be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt.
EX: "I know that I left the key on the table", "Galileo knew that the earth moves around the sun".

4) Be familiar or acquainted with a person or an object.
EX: "She doesn't know this composer", "Do you know my sister?", "We know this movie", "I know him under a different name", "This flower is known as a Peruvian Lily".

5) Have firsthand knowledge of states, situations, emotions, or sensations.
EX: "I know the feeling!", "Have you ever known hunger?".

6) Accept (someone) to be what is claimed or accept his power and authority.
EX: "The Crown Prince was acknowledged as the true heir to the throne".

7) Have fixed in the mind.
EX: "I know Latin", "This student knows her irregular verbs", "Do you know the poem well enough to recite it?".

8) Have sexual intercourse with.
EX: "Adam knew Eve".

9) Know the nature or character of.
EX: "We all knew her as a big show-off".

10) Be able to distinguish, recognize as being different.
EX: "The child knows right from wrong".

11) Perceive as familiar.
EX: "I know this voice!".

In this case, the senses apply to situations that other senses **can define**. In other words, the senses **overlap**. Examples of overlapping senses can be 1 and 5, since both imply the fact of knowing something, without stating a clear difference between them. Another example are senses 4 and 11, which both refer to be familiar with something, where 11 is more general than 4 and, therefore, could also apply to the situations 11 refers to. Is is also interesting the fact that these senses also share similar or exact words between them, which means that it seems that the main cause of senses to overlap is, in fact, the **similarity** between their definitions.

Also, the examples do not help distinguish the senses, since they are also **ambiguous**, in the sense that the examples provided can apply to other senses and, thus, **create more confusion** between senses. An example of this can be found in senses 1 and 3:

- **Sense 1 definition:** Be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about.
- **Sense 3 definition:** Be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt.
- **Example:** "I know that the President lied to the people" (Sense 1?).
- **Example:** "I know that I left the key on the table" (Sense 3?).

Without any further context, both sentences could apply to either one sense or another, since both could refer either to the simple knowledge or a fact or to strongly believe that something is true.

As it has been seen in these observations, and has been pointed out previously by López de Lacalle and Agirre (2015b) and Hovy et al. (2006), the **sense definitions** of a word seem to show characteristics that affect its difficulty when disambiguating it in context and are, indeed, ***too fine-grained*** in some cases and, as it has been observed, too **similar**.

1.2.2 Observations on the difficulty of context sentences and examples

Regarding the **context sentence**, it can be considered either *easy* or *difficult* by taking into consideration whether **the context provides enough information** to be able to choose between a sense or another, as López de Lacalle and Agirre (2015b) pointed out in

their work.

Thus, on the one hand, an ***easy*** context sentence **will contain enough information** to be disambiguated:

Word: *wait-v*

Senses:

- **1:** Stay in one place and anticipate or expect something.
- **2:** Wait before acting.
- **3:** Look forward to the probable occurrence of.
- **4:** Serve as a waiter or waitress in a restaurant.

Sentences:

- **1:** "Prudie, imagining herself **waiting** on *tables*, concurs that an appreciative gratuity is, indeed, preferable to repeated thank yous and now considers the problem solved". - **Sense 4**
- **2:** "Saatchi, by contrast, has kept a whole generation of artists from having to **wait** *tables*". - **Sense 4**

In this case, both examples belong to **Sense 4**. They are **easy**, since the sentence provides all the **necessary** information to distinguish the intended sense of the word.

In the example, there are **specific words** that automatically discard all the other senses, which are marked in italics. In general, these words are closely related to the sense: *table*, which is related to *work as a waiter/waitress*.

On the other hand, a ***difficult*** context sentence **will not contain enough information** for the word to be disambiguated, i.e. it is **underspecified**. The following examples are also from ***wait-v***:

- **1:** "Good things come to those who **wait**".
- **2:** "You'll have to **wait** until 9:30".
- **3:** "I **waited** until Jueli came to leave".

These sentences have something in common: they are **too general** in meaning. That is, they do not provide **enough** context to be able to discard the rest of senses. In fact, these sentences could apply to **nearly all the senses** of the word. For example, **sentence 2** can be interpreted according to nearly all the senses as follows:

- **Sense 1:** The hearer has to wait until 9:30 and something will happen at that time.

- **Sense 2:** The hearer has to wait until 9:30 to do something.
- **Sense 3:** Does not apply.
- **Sense 4:** The hearer has to work as a waiter/waitress until 9:30.

These observations have proven that context sentences and examples also show characteristics that make their interpretation difficult. As it has also been pointed out in López de Lacalle and Agirre (2015b), the lack of **specificity** of a context sentence can represent a problem when disambiguating.

1.3 Hypotheses

After having analysed the situations related to WordNet and word sense disambiguation, it is clear that words and context sentences show behaviours that may cause problems in a disambiguation task.

In this study, the experiments are based in the previous observations (the **ambiguity** in the senses and the **specificity** of the context sentences) so as to establish the factors that affect the difficulty of disambiguation of a word in context and up to what extent they do so, in order to be able to, first, detect difficult target words/context sentences relying on annotation data (supervised analysis) and, then, **predict** them without it (unsupervised methods), which is the final objective of this project.

In order to get to be able to predict the difficulty of words and sentences without annotation data, two initial hypotheses have been formulated:

- It is possible to predict the difficulty of a target word **without relying on annotation data** in terms of **similarity** between all the pairs of senses of said word.
- It is possible to predict the difficulty of a context sentence **without relying on annotation data** in terms of its **probability** and **length**.

1.4 Objective of this thesis

In this thesis, there are 3 main objectives:

- To be able to model **target word difficulty**, by using unsupervised methods, by taking into account the similarity between its senses, as stated in the **first hypothesis**.
- To be able to model **context sentence difficulty**, by using unsupervised methods, by analysing the **specificity** (that is, the amount of information that it is provided) of the context sentence, as stated in the **second hypothesis**.

- To be able to model **example** difficulty, by combining the two proposed unsupervised measures.

1.5 Thesis organization

.

In order to achieve the objectives, this thesis will be organized as follows:

- Section 2 will make reference to previous work on the field and important resources and contributions that will be used in the project.
- In Section 3, the supervised metrics to estimate the difficulty of target words and context sentences, kappa and entropy, respectively, will be presented.
- In Section 4, the first and second objectives will be addressed, where the unsupervised proposed metrics to estimate difficulty of target words and context sentences −similarity between sense definitions and probability, respectively−, will be presented. These values will be correlated to their analogous supervised metrics in order to prove the validity of each factor to measure difficulty.
- In Section 5, which addresses the third and last objective, a model to predict example difficulty will be designed, by using linear regression and the difficulty metrics obtained in Section 4.
- Section 6 will include the final conclusions and contributions of this thesis, along with further work that can be performed in the future.

2 State-of-the-art

The bibliography regarding this specific topic is scarce and related to WSD systems, although there are some works that are worth mentioning.

2.1 Resources: WordNet

One of the major sense inventories is **WordNet** (Miller, Beckwith, et al. 1990). WordNet is an on-line lexical database for English created at the University of Princeton in 1990. In it, **nouns**, **adjectives**, **verbs** and **adverbs** −and their senses− are stored and organized according to their semantic information into **synsets** or, as Fellbaum (2005) describes them, "sets of cognitive synonyms". In this sense, words are interlinked between them according to their senses and their semantic characteristics.

For **nouns**, the most remarkable relations are the ones related to synonymy (how two words have the same meaning, such as *car* and *automobile*) and the ones related to "super-subordinate relations" (Fellbaum 2005). Super-subordinate relations include "hyperonymy, hyponymy and ISA relations" (Fellbaum 2005).

Hyperonymy relations take into account how a word (**hyponym**) belongs semantically to a more general group (**hyperonym**). For example:

flower: {rose, lily, sunflower}

In this case, *flower* would be the **hyperonym** and *rose*, *lily* and *sunflower* would be the **hyponyms**. **ISA** relations (named after "is a") would be the relations between hyponyms and their hyperonyms. In the previous example, the relation would be as follows:

{rose, lily, sunflower} **ISA** flower

Another important relation between nouns stored in WordNet is **meronymy** and **holonymy**. **Meronymy** stands for the relations between a word (the **holonym**) and the words that refer to the **parts** of said word (that is, a **part-whole** relation). For example:

body: {arm, leg, head}

In the example, *arm*, *leg* and *head* are parts of the *body* and, therefore, are **meronyms** of *body*, the **holonym**. Also, the meronyms of a word will be inherited from the hyperonyms of said word, but not vice versa, since hyponyms have characteristics that make them different from their hyperonyms. A good example, included in Fellbaum (2005), is **furniture**:

Hyperonymy - hyponymy
furniture - chair, bed
chair - armchair
bed - bunkbed

Holonymy - meronymy
chair - legs
armchair - legs, arms

As it can be seen, *armchair* inherits the meronym *legs*, but *chair* cannot inherit *arms* from armchair, since *arms* are a characteristic that differentiates a *chair* from an *armchair*.

Regarding **verbs**, they are also organized into a hierarchy based on **troponymy**. The troponyms of a verb express a more specific way of performing the action designated by the verb. For example:

{communicate} - {talk} - {whisper} (Fellbaum, 2005)

In the example, *talk* is a specific way of *communicating* (by using speech), and *whisper* is a specific way of *talking*, in which a very soft voice is used.

Verbs are also related to other nouns in terms of semantic roles. For example:

{**paint**} - {picture} [RESULT] - {painter} [AGENT] (Fellbaum, 2005)

As it can be seen, a *picture* is the RESULT of the action of *paint*ing, whereas the *painter* is the AGENT (or "doer") of the action of *paint*.

Regarding **adjectives**, they are organized in terms of synonymy/antonymy (semantically similar and semantically opposite, respectively). Furthermore, in WordNet there are also **pertainymy** relations between adjectives and nouns. More specifically, *pertainmy* relates adjectives and the nouns they are derived from. For example:

{criminal} - {crime} (Fellbaum, 2005)

According to Fellbaum (2005), since most English adverbs are "derived from adjectives [...] (*surprisingly*, *strangely*, etc.)", there are not many adverbs in WordNet (*mostly*, *really*, etc.).

The analysis of difficulty of words in this project will use WordNet 3.0 word senses, as in Passonneau and Carpenter (2014).

2.2 Resources: MASC dataset

Passonneau and Carpenter (2012) performed an analysis of the results stored in the Manually Annotated SubCorpus (MASC), which is a corpus originated by the need of a high quality representation of linguistic phenomena. In 2010, Ide et al. (2010) extracted from the Open American National Corpus (OANC) -which includes 15 million words- a sample of a half million words in order to create MASC.

This corpus includes a wide variety of words, from different sources:

Genre	No. files	No. words	Pct corpus
Court transcript	2	30052	6%
Debate transcript	2	32325	6%
Email	78	27642	6%
Essay	7	25590	5%
Fiction	5	31518	6%
Gov't documents	5	24578	5%
Journal	10	25635	5%
Letters	40	23325	5%
Newspaper	41	23545	5%
Non-fiction	4	25182	5%
Spoken	11	25783	5%
Technical	8	27895	6%
Travel guides	7	26708	5%
Twitter	2	24180	5%
Blog	21	28199	6%
Ficlets	5	26299	5%
Movie script	2	28240	6%
Spam	110	23490	5%
Jokes	16	26582	5%
TOTAL	376	506768	

Figure 1: Distribution of the different data in the MASC corpus (Passonneau, Baker, et al. 2012)

Also, these words were analysed regarding different phenomena:

Annotation type	No. words
Logical	506659
Token	506659
Sentence	506659
POS/lemma (GATE)	506659
POS (Penn)	506659
Noun chunks	506659
Verb chunks	506659
Named Entities	506659
FrameNet	39160
Penn Treebank	*506659
PropBank	55599
Opinion	51243
TimeBank	*55599
Committed Belief	4614
Event	4614
Dependency treebank	5434

* under development

Figure 2: Types of analysis in the MASC corpus (Passonneau, Baker, et al. 2012)

In addition, and for the interest of this project, a small set of MASC corpus is also sense-tagged (with WordNet 3.0 sense labels) for 116 words, with approximately 1000 context sentences for each word, as Passonneau and Carpenter (2014) state. In their article, the authors created a subset of 45 words (17 nouns, 16 verbs and 9 adjectives), randomly

selected from the sense-tagged set. For each word, as mentioned previously, there were 1000 sentences and every sentence was sense-annotated by 20-25 different annotators by using a crowdsourcing platform (Amazon Mechanical Turk). It is interesting to point out that the annotators (or Turkers), were not trained nor familiar with WordNet.

In order to associate a single sense to each example, by using the crowdsourced data, the authors, instead of using a conventional model of annotation based on the number of annotations, such as **Most Frequent Sense** (MFS), present a new model, a probabilistic one. The model they propose is based on **Dawid and Skene**'s model (Dawid and Skene 1979).

MFS is a model of estimation that takes into account the observations that are observed more frequently in order to determine, in this context, the "correct"[5] sense. In a nutshell, thus, MFS will account as "correct" or "true", the **most annotated sense**.

The model the authors proposed, based on **Dawid and Skene**'s model (Dawid and Skene 1979), is a probabilistic model that is more complex, since it takes into account more factors in order to assign a single sense to an example by using annotations (Passonneau and Carpenter 2014).

This model takes into account four elements from the annotation data: the **annotator** (as the individual that annotates each example), the **annotations** (labels, represented as y), the **inventory of senses** of the word to disambiguate (Z), and the **examples** (word instances). The model uses this data in order to calculate three components:

- The **true category** of the word (z_i). That is, the *true* sense of the word. Note that the examples in the task do not have an assigned *correct* sense and it needs to be calculated probabilistically.

- The **accuracy and biases** from the annotator (θ). As its name indicates, it is the accuracy associated to the annotator given the estimated true categories.
  θ can be represented with a matrix like the following:

$$\theta_x = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \tag{1}$$

  In this matrix, the annotator x annotates a word with two senses. Each part of the matrix represents different probabilities. In this matrix:

  - $p(k' = 1|k = 1) = 0.9$. In other words, 0.9 is the probability that annotator x annotates sense 1 (k') when the *true* sense (k) is 1.
  - $p(k' = 2|k = 1) = 0.1$. In other words, 0.1 is the probability that annotator x annotates sense 2 (k') when the *true* sense (k) is 1.

  The rest of probabilities work in the same way.

[5]There is no mention of *correctness* in the annotation, due to the fact that the examples in the task do not have a previously assigned *correct* sense, in a strict sense of the word. Thus, the mention of *correct* will be between quotation marks.

- The **prevalence** for each sense of a word ($\pi_k, where\ k \in K$[6]). In a nutshell, the prevalence is the *a priori* probability of the senses.

It is possible to obtain the *true* sense of each example by using the elements described above and Bayesian estimation methods (Passonneau and Carpenter 2014).

The formula in order to estimate the *true* sense is the following:

$$\begin{aligned} p(z_i|y,\theta,\pi) &\propto p(z_i|\pi)p(y|z_i,\theta) \\ &= \pi_{z[i]} \prod_{ii[n]=i} \theta_{jj[n],z[i],y[n]} \end{aligned} \tag{2}$$

Where the variable *ii[n]* is the example index (i) of annotation n, annotated by an annotator *jj[n]* with an annotation *y[n]*. Given a *true* sense $z[i]$, $\pi_{z[i]}$ is the prevalence of $z[i]$ in the example (i). Finally, $\theta_{jj[n],z[i],y[n]}$ is the accuracy/bias of annotator *jj[n]*. In other words, it is the probability for annotator *jj[n]* to annotate *y[n]* given that the *true* category is $z[i]$.

The following example, extracted from (Passonneau and Carpenter 2014) is an example of how this rule is applied in order to estimate the *true* sense of an example. Suppose there is an annotation task in which three annotators have to annotate the same example for a word with two senses. Each annotator annotates an example with a label, and has accuracies/biases associated:

- $K = 2$. The word has two senses.
- $\pi_1 = 0.2$, $\pi_2 = 0.8$. Sense 1 has a prevalence of 0.2, whereas sense 2 has a prevalence of 0.8.
- Accuracies/biases:
  $\theta_1 = \begin{bmatrix} 0.75 & 0.25 \\ 0.40 & 0.60 \end{bmatrix}$ $\theta_2 = \begin{bmatrix} 0.65 & 0.35 \\ 0.30 & 0.70 \end{bmatrix}$ $\theta_3 = \begin{bmatrix} 0.90 & 0.10 \\ 0.20 & 0.80 \end{bmatrix}$
- Labels supplied for each annotator for example z_i: $y_1 = 1$, $y_2 = 1$, $y_3 = 2$

The application of the formula would be the following:

To estimate the probability of sense 1 to be the *true* sense:

$$Pr[z_i = 1|y,\theta,\pi] \propto \pi_1 \cdot \theta_{1,1,1} \cdot \theta_{2,1,1} \cdot \theta_{3,1,2} \tag{3}$$

$$Pr[z_i = 1|y,\theta,\pi] \propto 0.2 \cdot 0.75 \cdot 0.65 \cdot 0.1 = 0.00975 \tag{4}$$

[6]The inventory of senses of the word.

To estimate the probability of sense 2 to be the *true* sense:

$$Pr[z_i = 2|y, \theta, \pi] \propto \pi_2 \cdot \theta_{1,2,1} \cdot \theta_{2,2,1} \cdot \theta_{3,2,2} \tag{5}$$

$$Pr[z_i = 1|y, \theta, \pi] \propto 0.8 \cdot 0.4 \cdot 0.3 \cdot 0.8 = 0.0768 \tag{6}$$

After normalizing and rounding:

For sense 1:

$$Pr[z_i = 1|y, \theta, \pi] = \frac{0.00975}{0.00975 + 0.0768} = 0.11 \tag{7}$$

For sense 2:

$$Pr[z_i = 2|y, \theta, \pi] = \frac{0.0768}{0.00975 + 0.0768} = 0.89 \tag{8}$$

In this case, the probability of **sense 1** to be the *true* sense is **0.11**, whereas the probability of **sense 2** to be the *true* sense is **0.89**. Therefore, the *true* sense estimated by the model in this example would be **sense 2**.

The experiments in this thesis will be based in the results of this task.

2.3 Factors of difficulty

Linguistically-motivated factors

In the field of Linguistics, Koirala and Jee (2015) tried to define the factors that had more impact in the gradience of sentence difficulty. They distinguished two types of features:

- **Traditional features (0, 1)**
  - **Number of words**. Number of words of the sentence.
  - **Low-frequency words**. Number of words with low frequency (that is, words that are not very common or are more specific).
- **Non-traditional features (2-7)**
  - **Counts of clauses**. Number of constructions with a subject and a verb.
  - **Dependent clauses**. A dependent clause differs from a regular clause in the fact that it cannot stand alone as a single entity (it **depends** of another sentence).
  - **Coordinate phrases**. Coordinated elements, joined by a coordinate conjunction, such as *and* or *or*.
  - **T-units**. A T-unit is "one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it" (Lu 2010).
  - **Complex t-units**. A complex T-unit "contains a dependent clause" (Lu 2010).

– **Wh nominals**. Wh-elements (*who*, *what*, *why*, etc.) used in non-question contexts, such as *who* in "I know *who* the murderer is".

The procedure consisted in surveying some subjects, which had to classify the difficulty of the sentences they were provided in a scale from 1 to 4, being 4 the maximum difficulty. With these results, the authors could state that, for both traditional and non-traditional features, the difficulty increased when the values for the features increased.

With those results, the authors tried to estimate the importance of each factor in determining the difficulty of a sentence by using a *random forest* classifier.

The results obtained were the following:

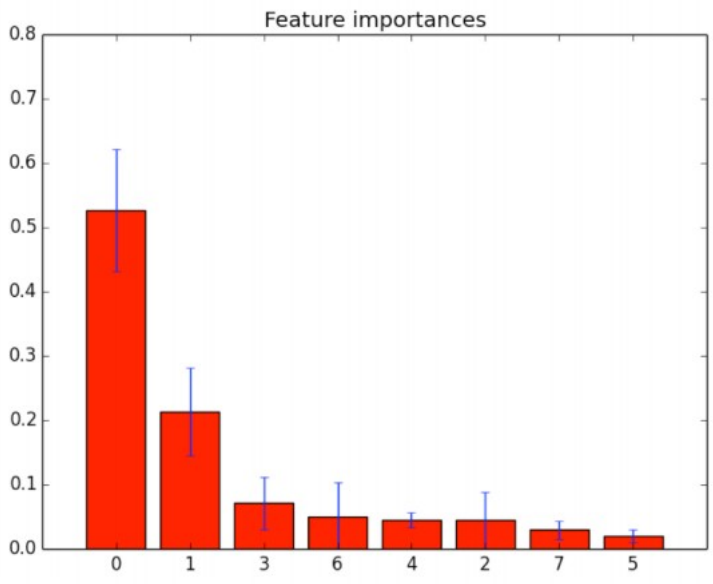


Figure 3: Factors that influence sentence difficulty. Results from Koirala and Jee (2015)

Figure 3 shows a bar plot in which the importance for each factor is indicated. Each number corresponds to a factor: 0 corresponds to number of words, 1 corresponds to number of low-frequency words, and 2 to 7 correspond to non-traditional features.

Thus, it became proved that traditional features (length and number of low-probability words) are key in order to determine the difficulty of a sentence. These features will serve as an inspiration to model context sentence difficulty in the analysis performed in the next sections.

Miller and Leacock (2000) try to describe the state-of-the-art in relation to disambiguation of polysemous words (that is, words with more than one meaning). The authors remark the importance of the context, since it provides "semantic as well as syntactic information" (Miller and Leacock 2000) and, therefore, provides information that may determine the meaning of the word and its form, which can be very important information in the task of disambiguating polysemous words.

Computationally-motivated factors

Yarowsky and Florian (2002) analyse the performance of various WSD systems with a set of factors:

> (a) target language (English, Spanish, Swedish and Basque); (b) part of speech; (c) sense granularity; (d) inclusion and exclusion of major feature classes; (e) variable context width (further broken down by part-of-speech of keyword); (f) number of training examples; (g) baseline probability of the most likely sense; (h) sense distributional entropy; (i) number of senses per keyword; (j) divergence between training and test data; (k) degree of (artificially introduced) noise in the training data; (l) the effectiveness of an algorithm's confidence rankings; and (m) a full keyword breakdown of the performance of each algorithm. (Yarowsky and Florian 2002)

In order to be able to analyse each feature, the authors perform several experiments, and compare the results by using different classifiers(Yarowsky and Florian 2002):

- Cosine vector model (Cosine)
- Non-hierarchical decision lists (DL)(Yarowsky and Florian 2002)
- Transformation-Based Learning (TBL). More specifically, a variant for WSD (Yarowsky and Florian 2002)
- Naïve Bayes (NaiveBayes)
- BayesRatio model(BR)
- Feature-Enhanced Naïve Bayes (FENBayes)
- Majority sense

The authors tried to analyse the interaction between the factors and the performance of the classifiers. The experiments, among others, included (Yarowsky and Florian 2002):

- Analysing the accuracy of the classifiers according to language.
- The effect of the context sentence window size (that is, length) on accuracy.
- Measure the performance of the system according to the number of training examples.
- Analysis of the number of senses of the target word.

- Measure of the probability that has the most probable sense.
- Measure the sense entropy.
- Measure the effect of noise in the annotations on accuracy.

For each experiment, the following conclusions can be extracted (Yarowsky and Florian 2002):

- There is not a *best classifier* for all languages. Depending on the language, there is a classifier that achieves better accuracy. For example, in Spanish and Swedish, cosine, FENBayes and BR perform better, but in English and Basque cosine performs the worst.
- The context window size affects accuracy depending on the classifier. In general, smaller windows achieve better results, and when the window is +100, the results remain constant (in high values). Also, the authors observed that, for adjectives, the results were more variable and different among classifiers.
- The more training examples for each sense, the better results and, therefore, the less error.
- The more number of senses per target word, the worse results.
- The higher the probability of the majority sense, the better results.
- The higher the sense entropy, the worse results.
- The more noise, the worse results.

The results of their research, thus, provided very interesting data about factors that may cause an example to be difficult. Taking into account that the authors performed the analysis by using annotated data, this thesis will go one step forward: instead of only trying to find supervised factors to measure difficulty, **also** equivalent metrics to measure difficulty on unannotated data will be proposed and applied.

In their article, Martínez Alonso et al. (2015) state that there are some factors that make total agreement impossible. These factors may be related to the sense inventory, the examples to annotate or the annotators.

The main goal of their article is trying to predict agreement on word-annotation by analysing the linguistic properties of the example. In order to do so, they use **9** sense-annotated datasets (Martínez Alonso et al. 2015):

- **MASCC** - English crowdsourced sense-tagged corpus.
- **MASCE*** - Expert annotated corpus. It consists of **four** sub-corpora: MASCEW (all rounds together) + rounds 2, 3, 4.

- **FNTW** - English Twitter FrameNet corpus.
- **ENSST** - English supersense-annotated corpus.
- **EUSC** - Basque lexical-sample SemCor.
- **DASST** - Danish supersense-annotated data.

As it can be seen, the corpora could be expert- or crowdsource- annotated.

The authors take into account three elements in their analysis (Martínez Alonso et al. 2015):

- Sentence (s)
- Word (w)
- Syntactic parent (p)

Furthermore, for their analysis make use of 19 features, that represent the frequency (2 features), morphological (5 features), syntactic (5 features), contextual (5 features) and sense-inventory (2 features) characteristics of the examples to analyse (Martínez Alonso et al. 2015).

The results obtained can be summarized in three main points (Martínez Alonso et al. 2015):

- The **best results** have been obtained in datasets with **lots of annotations** and **more than five annotators**. Also, the **size** of the dataset is a relevant information.
- Three features obtained the **highest correlation with agreement**:
  - Sense entropy (sense inventory)
  - Number of labels (sense inventory)
  - Frequency of the target word (frequency)
- The **sense inventory** is a **very valuable** factor in order to be able to predict agreement.

The features used by the authors can be related to the experiments performed in this thesis, since the approach in order to perform the analysis will be linguistically-motivated.

Furthermore, López de Lacalle and Agirre (2015b) show how problematic examples cannot be used for training WSD systems, since they decrease the performance of said systems. Thus, the authors try to detect these problematic examples using crowdsourced annotation. More specifically, they use the data from Passonneau and Carpenter (2014)'s annotation

task, which is also the data that will be used in this thesis.

In order to detect problematic examples, the authors correlate three factors to the performance of the system (IMS - *It Makes Sense*) (Zhong and Ng 2010). These factors included two extracted from Yarowsky and Florian (2002): **number of senses of the word** and **sense entropy**. The third one, as an addition from the authors, was the **annotation entropy**. As it can be seen in Figure 4, in which the result of the correlations between performance and the factors (number of senses, sense entropy and annotation entropy) is indicated, the results proved that **annotation entropy** was the most correlated factor, being the **number of senses** the least correlated:

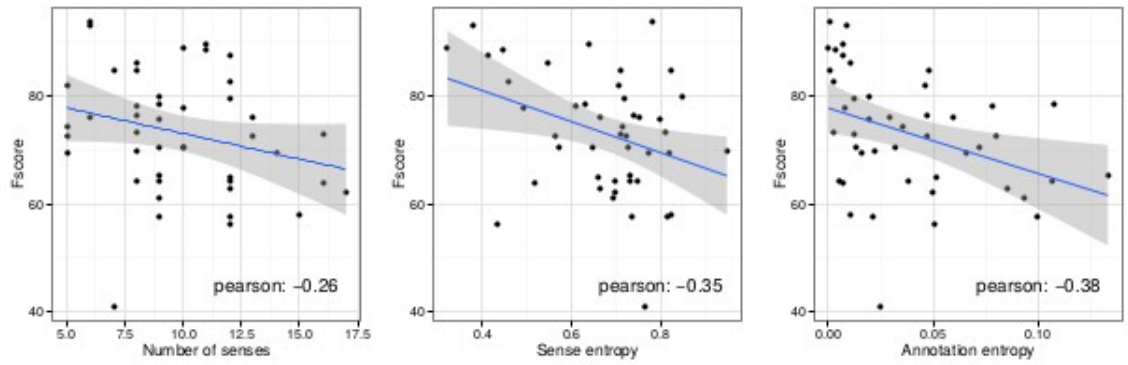


Figure 4: Correlation results, extracted from López de Lacalle and Agirre (2015b)

In order to prove the results, the authors evaluated the performance of the system by removing problematic examples in some words (the ones that had more entropic annotations). The results showed that these words had better results without the most entropic annotations than with all the annotations.

Also, they established two hypothetical factors that could affect the entropy of an example: the **confusion between senses** and **insufficient context**.

Furthermore, the study shows that the inter-annotator agreement shows a relationship with word performance. For this, they assume that the higher the confusion, the more difficult a word is.

The same authors, in another of their articles (López de Lacalle and Agirre 2015a), try to increase accuracy of WSD systems by clustering word senses. The authors claim that this system may increase accuracy to a 90%. The data they use comes from the crowdsourced annotation task in Passonneau and Carpenter (2014), tagged with WordNet 3.0 senses.

In the article, the authors assume that if the confusion between two senses is high, the senses are difficult to discriminate (López de Lacalle and Agirre 2015a). They also add that the context the target word is in may be another factor of difficulty, since it can be underspecified, but they leave this aspect aside in the article.

Thus, from this assumption, the authors create a confusion matrix for each target word, in order to create the clusters.

The authors distinguish between three different groupings for target words' senses (López de Lacalle and Agirre 2015a):

- Fine-grained senses. The original senses of the target word.
- Clustering of senses according to confusion. If the confusion is high, the senses tend to be in the same cluster.
- Random clustering.

In order to perform the experiments, the authors use as a gold standard the estimations obtained by Dawid and Skene's probabilistic model, obtained from Passonneau and Carpenter (2014).

The experiments were performed by using the clusters and the WSD algorithm *It Makes Sense* (IMS), from Zhong and Ng (2010), which had the best disambiguation results to date (López de Lacalle and Agirre 2015a). The results obtained by IMS were compared to the ones obtained with the Most Frequent Sense (MFS), estimated by using the training corpus (López de Lacalle and Agirre 2015a).

The bar plots in Figure 5 show the results grouped according to each sense inventory (fine-grained, random clustering, and confusion clustering, respectively). The first plot corresponds to the accuracy obtained for each group and each algorithm, whereas the second shows the error reduction of IMS in relation to MFS (López de Lacalle and Agirre 2015a).

As it can be seen, the results show that the confusion clustering obtains very good results for accuracy (reaching 92.6%) and the error reduction of IMS in relation to MFS is considerable. Thus, the experiments performed by the authors show that clustering is a valid approach in order to improve WSD systems.

It is worth mentioning, too, that the clusters often consist of similar senses, as Table 1 shows.

Conf	**Gloss**
1	a relative position or degree of value in a graded group
1	a specific identifiable position in a continuum or series or especially in a process
1	a position on a scale of intensity or amount or quality
2	height above ground
5	an abstract place usually conceived as having depth
6	a structure consisting of a room or set of rooms at a single position along a vertical scale
4	a flat surface at right angles to a plumb line
3	indicator that establishes the horizontal when a bubble is entered in a tube of liquid

Table 1: Relation of clustered senses by confusion (`Conf`) and original WordNet 3.0 sense definitions (`Gloss`) for ***level-n***, adapted from (López de Lacalle and Agirre 2015a)

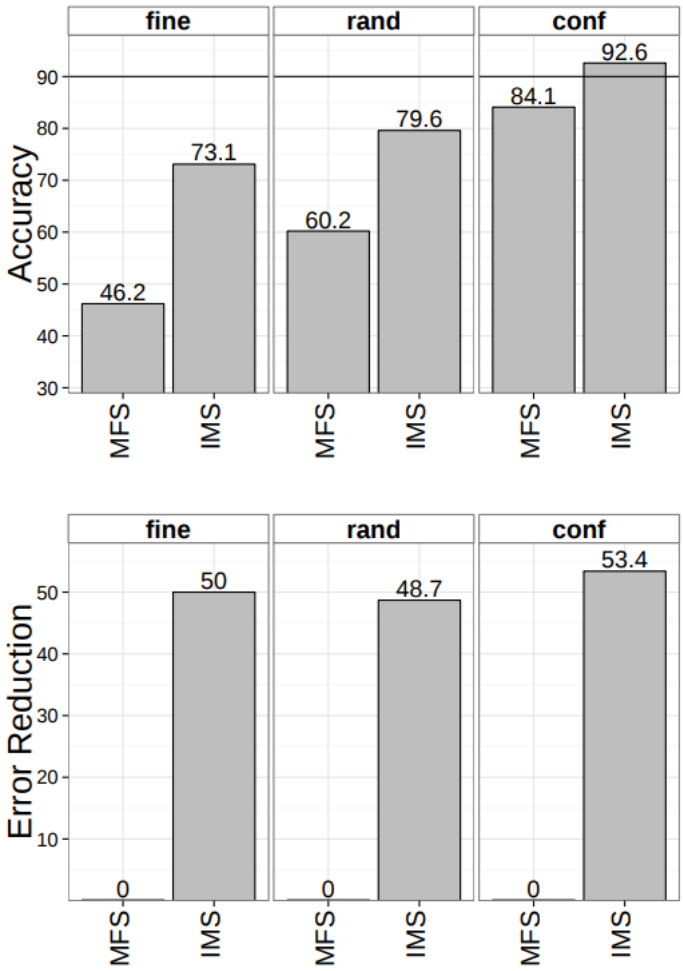


Figure 5: Accuracy of sense inventories and algorithms, and error reduction of IMS in relation to MFS, extracted from (López de Lacalle and Agirre 2015a)

Table 1, adapted from (López de Lacalle and Agirre 2015a), shows the clusterings obtained for the word ***level-n*** from the confusion clustering algorithm designed by the authors, and the WordNet senses for each of the clusters. As it can be observed, the senses in cluster 1 express similar ideas, and even share the same words, or synonymous ones (*position*, *degree-scale*...).

These observations are valuable information in order to analyse the problems in difficulty regarding target words and their sense definitions. From the studies in this section, we can extract which are the two factors to take into account when trying to predict the difficulty of disambiguation of words in context: the WordNet senses for that word, and the context sentences they are in, as it will be seen in the next sections.

3 Estimation of difficulty using annotated data

As it has been mentioned in the previous section, there are two factors that seem to be related to the difficulty of disambiguation of a target word in context: the **similarity between its sense definitions**, and the **specificity of the context sentence**. In this case, the data stored in a corpus will be used to predict this difficulty.

3.1 Data

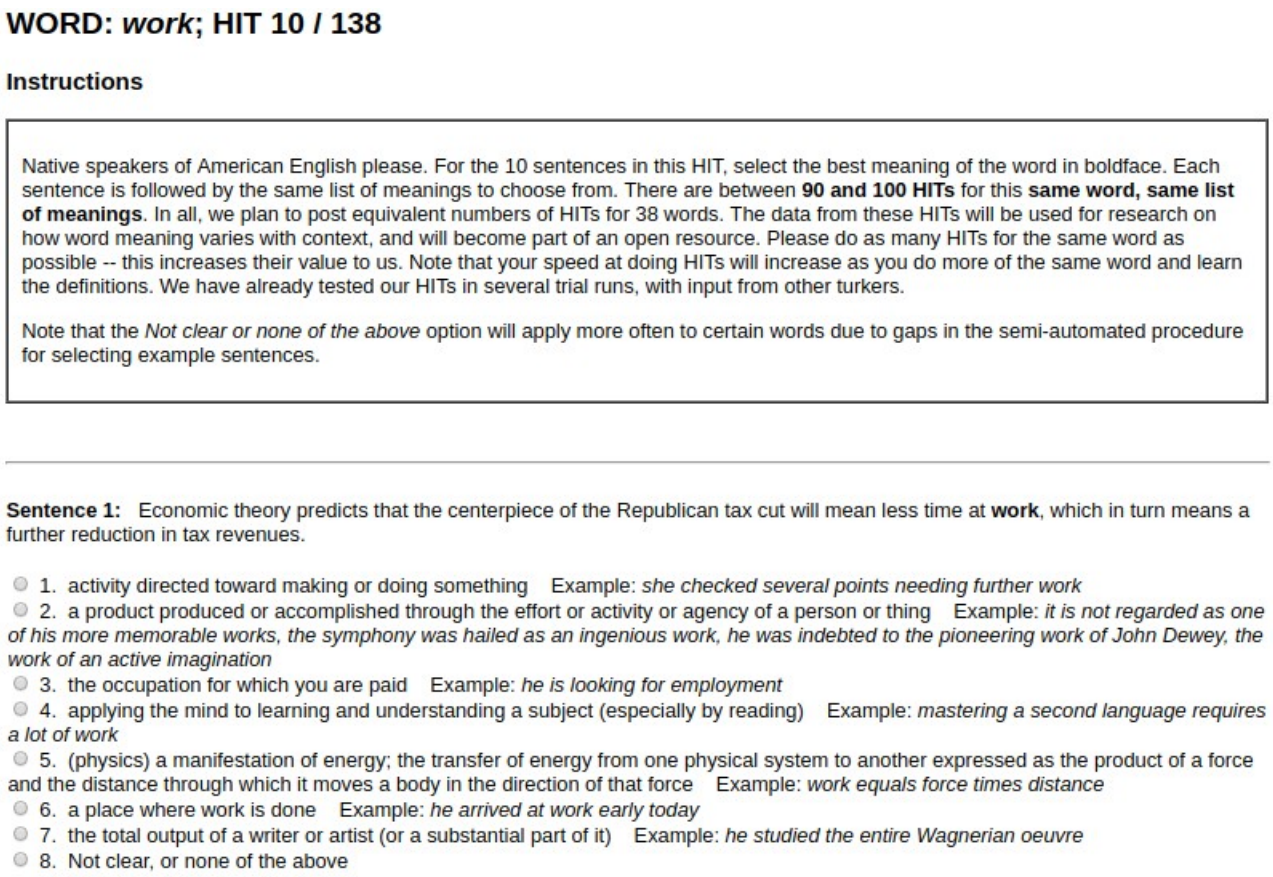
The data that will be used in order to estimate the difficulty of disambiguation will come from the annotation task in Passonneau and Carpenter (2014), described in 2.2. Therefore, the data will consist of 45 words (17 nouns, 16 verbs and 9 adjectives) and around 1000 context sentences per each word. These context sentences are annotated by 20-25 annotators each, by using WordNet 3.0 senses.

The words included in the corpus are the following:

- add-v
- appear-v
- ask-v
- board-n
- book-n
- color-n
- common-j
- control-n
- date-n
- fair-j
- family-n
- find-v
- fold-v
- full-j
- help-v
- high-j
- image-n
- kill-v
- know-v
- land-n
- late-j
- level-n
- life-n
- live-v
- long-j
- lose-v
- meet-v
- normal-j
- number-n
- paper-n
- particular-j
- poor-j
- read-v
- say-v
- sense-n
- serve-v
- show-v
- suggest-v
- tell-v
- time-n
- wait-v
- way-n
- win-v
- window-n
- work-n

Figure 6 shows an example of the task, which included, firstly, a set of instructions, which also included some details about the task itself (how many examples -HITs- per word are in the task, how many different words will compose it, and its purpose). The volunteers were also provided a sentence with a word in bold and its set of WordNet senses (with an example sentence for each of them), with an extra sense[7], and were demanded to disambiguate the **example** by using one of the senses provided.

[7]The extra sense is added to all words in order to account for the context sentences that the annotator considers that cannot be disambiguated by using the other senses.



WORD: *work*; HIT 10 / 138

Instructions

Native speakers of American English please. For the 10 sentences in this HIT, select the best meaning of the word in boldface. Each sentence is followed by the same list of meanings to choose from. There are between **90 and 100 HITs** for this **same word, same list of meanings**. In all, we plan to post equivalent numbers of HITs for 38 words. The data from these HITs will be used for research on how word meaning varies with context, and will become part of an open resource. Please do as many HITs for the same word as possible -- this increases their value to us. Note that your speed at doing HITs will increase as you do more of the same word and learn the definitions. We have already tested our HITs in several trial runs, with input from other turkers.

Note that the *Not clear or none of the above* option will apply more often to certain words due to gaps in the semi-automated procedure for selecting example sentences.

Sentence 1: Economic theory predicts that the centerpiece of the Republican tax cut will mean less time at **work**, which in turn means a further reduction in tax revenues.

1. activity directed toward making or doing something Example: *she checked several points needing further work*
2. a product produced or accomplished through the effort or activity or agency of a person or thing Example: *it is not regarded as one of his more memorable works, the symphony was hailed as an ingenious work, he was indebted to the pioneering work of John Dewey, the work of an active imagination*
3. the occupation for which you are paid Example: *he is looking for employment*
4. applying the mind to learning and understanding a subject (especially by reading) Example: *mastering a second language requires a lot of work*
5. (physics) a manifestation of energy; the transfer of energy from one physical system to another expressed as the product of a force and the distance through which it moves a body in the direction of that force Example: *work equals force times distance*
6. a place where work is done Example: *he arrived at work early today*
7. the total output of a writer or artist (or a substantial part of it) Example: *he studied the entire Wagnerian oeuvre*
8. Not clear, or none of the above

Figure 6: Example of task for word ***work-n*** (from Passonneau and Carpenter (2014)'s annotation task)

In this case, for example, the sense assigned would be **6**.

Furthermore, the estimated senses for the annotation model based on Dawid and Skene (1979) model are also provided.

By using the data stored from these annotations, the difficulty of a word and a context sentence can be estimated, as it will be shown in the next sections.

3.2 Estimating word difficulty

As it has been described in Section 2.2, Passonneau and Carpenter (2014) present a new model of annotation, based on probabilistic methods, which takes into account the **performance of the annotators** in the annotation task, based on Dawid and Skene (1979)'s model. This new model contrasts with the conventional models such as **MFS**, which, as mentioned in Section 2.2, takes into account the most annotated sense for each example as ground truth.

It is not easy to decide which model to use, since both appear to be useful. On the

one hand, the **MFS** model is simpler and more intuitive when interpreting the results, whereas Dawid and Skene (1979)'s model, although is more complex, takes into account more factors in order to estimate the probabilities of senses.

Table 2 shows a sample of the senses obtained for each annotation model: Dawid and Skene (1979)'s model (`D&S Sense`), and **MFS**. Also, the ID of the item (example) and the number of annotators for each item are provided:

	Word	ItemID	#Annotators	D&S Sense	MFS
1	add-v	1	25	6	1
2	add-v	10	25	6	1
3	add-v	100	25	1	1
4	add-v	101	25	3	1
5	add-v	102	25	2	2
6	add-v	103	25	6	1
7	add-v	104	25	1	1
8	add-v	105	25	3	3
9	add-v	106	25	1	1
10	add-v	107	25	1	1

Table 2: Sample of annotations obtained for each model of annotation, for some examples from ***add-v***.

As it can be seen in Table 2, in some cases the models do not agree. Figure 7 shows an example of this disagreement.

Figure 7 is a mosaic plot [8], in which the two models are compared by showing their agreement. It displays the cell frequencies of a contingency table in which the area of the boxes of the plot are proportional to the cell frequencies of the contingency table. The left plot corresponds to the agreement for the word ***know-v***, whereas the right plot shows the agreement for the word ***sense-n***. By comparing both plots, it can be observed that ***know-v*** shows more disagreement between models than ***sense-n*** and, therefore, more variability in the annotations, what causes the word to be more difficult to annotate. For this reason, the assumption is that words with a **high disagreement** rate are **difficult** words. In this project, **kappa agreement** will be used to estimate the difficulty of a target word by using annotation data.

3.2.1 Kappa agreement

Kappa agreement (in this case, Cohen's kappa coefficient) is a measure to calculate inter-rater (in this case, inter-model) agreement, by taking into account both agreement between raters and the probability of rating by chance. As Cohen (1960) states, "it is the proportion of agreement after chance agreement is removed from consideration".

[8] A mosaic plot is a graphical representation that helps recognizing the relationship between variables.

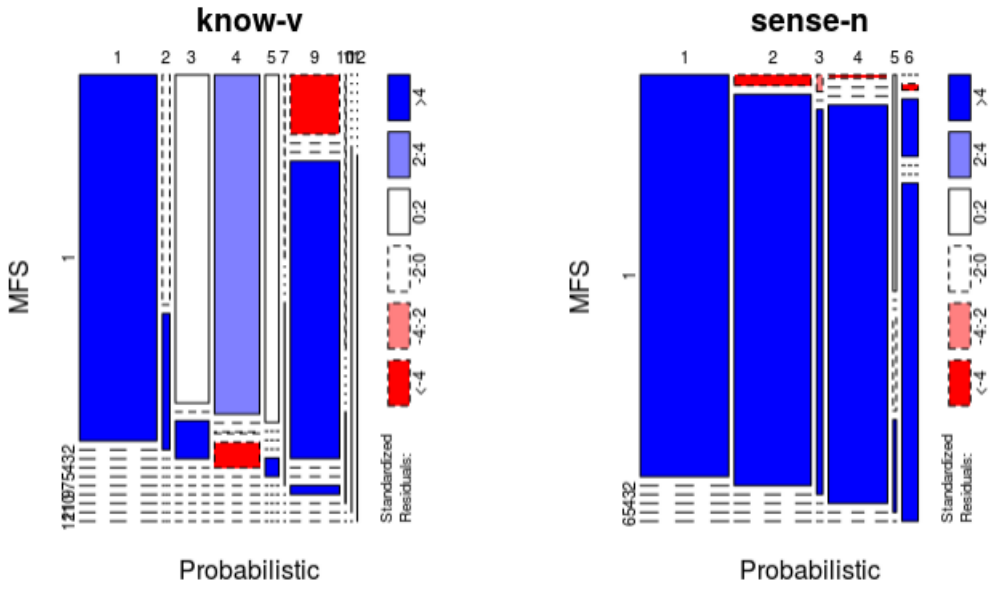


Figure 7: Mosaic plot for ***know-v*** and ***sense-n***

The formula to obtain the kappa value (Cohen 1960) is the following:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

Where (Cohen, 1960):

- p_o = the proportion of agreement between raters.
- p_e = probability of agreement by chance.

The formula will return a value between $[\frac{-p_e}{1-p_e}, 1]$, being $\frac{-p_e}{1-p_e}$ the least agreement[9] and 1 the maximum agreement.

For example, supposing that each model has estimated the senses for the same n examples of a word, and the target word has two senses, a contingency table can be created, in order to show the agreement and disagreement between the two models in regard to the two senses of the target word:

[9]The lowest agreement value depends on p_e.

		MFS	
		1	2
Probabilistic model	1	a	b
	2	c	d

Table 3: Annotation structure

In Table 3:

- The total number of estimated senses is $a+b+c+d$.
- The probabilistic model estimated **sense 1** $a + b$ times and **sense 2** $c + d$ times.
- MFS estimated **sense 1** $a + c$ times and **sense 2** $b + d$ times.
- The probabilistic model and MFS agreed on **sense 1** a times.
- The probabilistic model and MFS agreed on **sense 2** d times.
- The probabilistic model estimated **sense 1** when MFS estimated **sense 2** b times.
- The probabilistic model estimated **sense 2** when MFS estimated **sense 1** c times.

In order to obtain the elements necessary to calculate κ:

- To obtain p_o:

$$p_o = \frac{a+d}{a+b+c+d} \tag{10}$$

- In order to obtain p_e, the procedure is more complex.

  (1) First, it is necessary to determine the probability of the models both estimating sense 1 randomly:

$$p_1 = \frac{a+b}{a+b+c+d} * \frac{a+c}{a+b+c+d} \tag{11}$$

  Where the first operation stands for the probability of the probabilistic model estimating sense 1, and the second stands for the probability of MFS estimating sense 1.

  (2) After that, the probability of the models both estimating sense 2 randomly is calculated:

$$p_2 = \frac{c+d}{a+b+c+d} * \frac{b+d}{a+b+c+d} \tag{12}$$

  Where the first operation stands for the probability of the probabilistic model estimating sense 2, and the second stands for the probability of MFS estimating sense

2.

(3) The probability of estimating by chance will be the sum of both percentages:

$$p_e = p_1 + p_2 \tag{13}$$

A practical example

In this example, the target word has two senses, and the two models −the probabilistic model and MFS−, estimate the senses for 90 examples of the target word:

		MFS	
		1	2
Probabilistic model	1	30	15
	2	10	35

Table 4: Annotation example results

In this table:

- The total number of estimated senses is 90 (both models estimate the senses for the same 90 sentences).
- The probabilistic model estimated **sense 1** 45 times and **sense 2** 45 times.
- MFS estimated **sense 1** 40 times and **sense 2** 50 times.
- The probabilistic model and MFS agreed on **sense 1** 30 times.
- The probabilistic model and MFS agreed on **sense 2** 35 times.
- The probabilistic model estimated **sense 1** when MFS estimated **sense 2** 15 times.
- The probabilistic model estimated **sense 2** when MFS estimated **sense 1** 10 times.

Obtention of κ:

$$p_o = \frac{30+35}{30+15+10+35} = 0.72 \tag{14}$$

$$p_1 = \frac{30+15}{30+15+10+35} * \frac{30+35}{30+15+10+35} = 0.5 * 0.72 = 0.36 \tag{15}$$

$$p_2 = \frac{10+35}{30+15+10+35} * \frac{15+35}{30+15+10+35} = 0.5 * 0.55 = 0.275 \tag{16}$$

$$p_c = 0.36 + 0.275 = 0.635 \tag{17}$$

$$\kappa = \frac{0.72 - 0.635}{1 - 0.635} = 0.23 \tag{18}$$

Thus, for this annotation, κ would be 0.23, what would mean a low agreement between the probabilistic model and MFS.

3.2.2 Kappa agreement and word difficulty

Since kappa agreement value seems to be a robust value to show agreement, it will be considered as a measure to determine word difficulty.

Thus,

- A word is ***easy*** if the **agreement (kappa value)** for the annotations of said word is **high**.
- A word is ***difficult*** if the **agreement (kappa value)** for its annotations is **low**.

Therefore, the words in the dataset would be classified as follows, from less to more difficult or, in other words, from a higher Kappa value to lower Kappa value:

- sense-n - 0.96
- board-n - 0.93
- late-j - 0.92
- fair-j - 0.92
- read-v - 0.90
- live-v - 0.87
- time-n - 0.83
- ask-v - 0.83
- normal-j - 0.82
- life-n - 0.81
- common-j - 0.81
- meet-v - 0.78
- high-j - 0.77
- land-n - 0.77
- paper-n - 0.76
- long-j - 0.76
- poor-j - 0.75
- control-n - 0.73
- find-v - 0.73
- serve-v - 0.72
- lose-v - 0.72
- particular-j - 0.71
- tell-v - 0.69
- show-v - 0.69
- family-n - 0.68
- image-n - 0.68
- kill-v - 0.66
- full-j - 0.65
- number-n - 0.64
- add-v - 0.64

- window-n - 0.63
- work-n - 0.57
- way-n - 0.56
- appear-v - 0.56
- wait-v - 0.56
- suggest-v - 0.55
- date-n - 0.49
- say-v - 0.47
- win-v - 0.39
- help-v - 0.37
- fold-v - 0.37
- book-n - 0.35
- color-n - 0.26
- level-n - 0.24
- know-v - 0.13

In a graphical way:

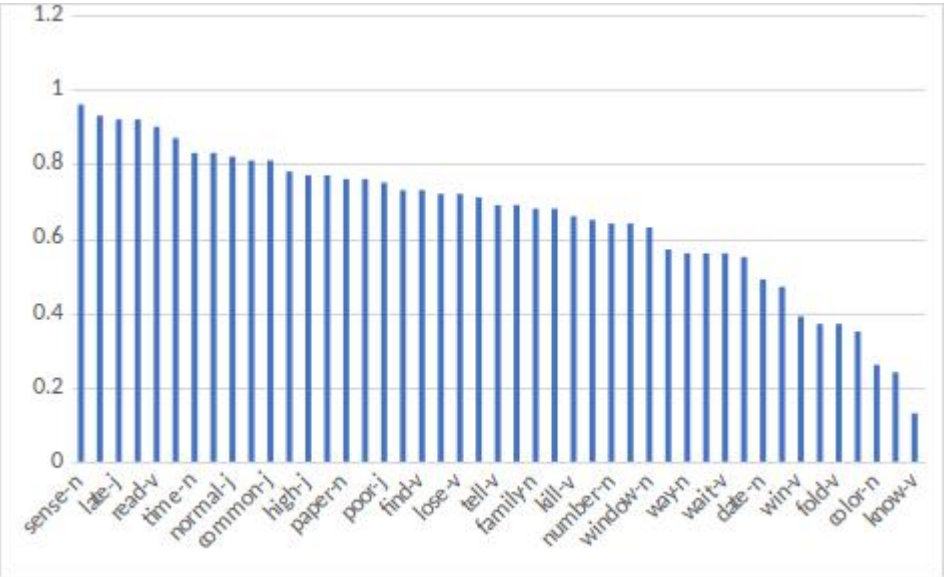


Figure 8: Classification of words according to their Kappa value

3.3 Estimating example difficulty

For each annotation example, and according to the annotations of that example by all the annotators, it is possible to calculate the entropy value for each one.

Entropy, unlike agreement, is a measure that takes into account the **variability** in the distribution of the probability of all the senses of a word to be annotated in an example. This means that it provides a value that accounts for the sparsity of the probabilities of each sense to be assigned to an example.

An intuitive way of understanding entropy would be the following:

	Sense 1	**Sense 2**	**Sense 3**	**Sense 4**
Example 1	0.5	0.3	0.2	0
Example 2	0.5	0.48	0.01	0.01

Table 5: Probabilities of annotating each sense for a word in two different examples

Table 5 shows a fictitious case, in which the target word has four senses, and it has to be disambiguated in two different examples. For each example, the probability[10] of each sense to be the *true* sense is provided.

In both examples, the most probable sense is sense 1, with a probability of 0.5. In Example 1, the probabilities are distributed between senses 1, 2 and 3, with similar probabilities, whereas in Example 2 the probabilities are distributed between senses 1, 2, 3 and 4. However, the most probable senses in the latter are 1 and 2, which are consistently more probable than senses 3 and 4.

As it can be seen, in spite of having the same most probable sense, with the same probability, it is clear that Example 1 and Example 2 should not be considered as equal in terms of sense disambiguation. In fact, it can be complicated to choose which example would be more difficult to annotate, since **Example 1** has **a lower number of probable senses**, but the **probabilities are close**, whereas in **Example 2** the probabilities for sense 1 and sense 2 are so close that it seems to suggest that both senses are **almost equivalent** in that context.

For this, it is necessary to have a measure that accounts this variability in order to determine how difficult to annotate is an example, so as to have a value that is more specific than agreement: **entropy**.

3.3.1 Entropy value

Entropy is a measure for ***disorder*** (that is, the disparity) in the annotations of a given **example**, taking into account the probabilities of annotating each sense for the word in a context.

The formula to obtain entropy (López de Lacalle and Agirre 2015b) is the following:

$$Entropy_{example} = -\sum_{i=1}^{p} p(i) log_2 p(i) \tag{19}$$

Where:

- i = The sense of the word to disambiguate.
- p(i) = The probability of $sense_i$. The probability estimations of sense annotations are extracted from the estimations from the probabilistic model presented in Passonneau and Carpenter (2014).

[10]Extracted from the probabilistic model presented in Passonneau and Carpenter (2014).

In order to be able to compare the words in the dataset, which have a different number of senses, the entropy is normalized by the number of senses (López de Lacalle and Agirre 2015b). Therefore, the final formula in order to calculate the entropy in the annotating task is the following (López de Lacalle and Agirre 2015b):

$$Entropy_{annotation} = \frac{Entropy_{example}}{\#senses} \tag{20}$$

3.3.2 Entropy and example difficulty

Since entropy, in a broad sense, is a measure that models the ***disorder*** in the annotations of a given **example**, the assumption is the higher the disorder in the annotation of an example, the more difficult an example is and, equivalently, the lower the disorder in the annotation, the easier an example is.

Thus,

- An example is ***easy*** if the **disorder (entropy)** in the annotations is **low**.
- An example is ***difficult*** if the **disorder (entropy)** in the annotations is **high**.

3.3.3 Entropy as a measure to model context sentence difficulty

As it has been mentioned previously, **entropy** is a measure that models the difficulty of **examples** and, therefore, measures difficulty by taking into account the difficulty of both the **target word** and the **context sentence**. However, it is also possible to measure the difficulty of **context sentences** using entropy.

This assumption comes from the fact that the high entropy value of a difficult example may be caused by different scenarios:

- Easy target word + Difficult context sentence
- Difficult target word + Easy context sentence
- Difficult target word + Difficult context sentence

As it can be seen, it is very difficult to establish which is the cause of high entropy in difficult examples.

However, for **easy examples**, the intuition becomes clearer, since their low entropy value can only be caused by one scenario: that **the target word is *easy*** (*easy* target words generally have a lower overall entropy value than difficult words, as it will be seen in next section), and **the context sentence is *easy***. Following this intuition, and as it has been pointed out previously, if the target word is *easy* and the context sentence is *difficult*, then the entropy value should be higher. To sum up:

- An *easy* target word in an *easy* context sentence $\longrightarrow$ low entropy.
- An *easy* target word in a *difficult* context sentence $\longrightarrow$ high entropy.

Taking into account these considerations, since it seems to be possible to model difficulty of the context sentences of easy words by using entropy, this intuition can be extended to both easy and difficult words.

3.4 Relations between words and examples

In order to wrap up this chapter, the relationship between the supervised measures to model target word and example/context sentence difficulty (kappa agreement and entropy, respectively) will be analysed. The goal is to ensure that both metrics are correlated to some extent in order to use them for further experimental analysis in the next sections.

3.4.1 Motivation

The motivation for this analysis comes from the assumption that the difficulty of a word is related to its context sentences. In this sense, the examples of an ***easy* word** should have a **lower overall entropy** than the examples of a ***difficult* word**, which are expected to have a **higher overall entropy**.

3.4.2 Analysis

Figure 9 represents the relation between kappa agreement (in the X axis) and the distribution of entropy of a word (in the Y axis), represented by a boxplot for each word.

Also, the boxplots for each word are sorted by their kappa value, in ascending order. They are also coloured according to their kappa value: $\kappa > 0.08$, $0.8 > \kappa > 0.4$, and $\kappa < 0.4$, in red, green and blue, respectively.

For each word, the representation and analysis will be focused on two characteristics of the distribution of the entropy of its examples:

- **Data range**. It is represented by the **boxplot** [11] of each word. The assumption is that examples from easy words will lie in lower entropy ranges than examples from difficult words.
- **Most frequent value**. It can be obtained by using **measures of centrality**, such as the mean, the median and the mode. Since mean is not a robust measure and mode is complex to obtain due to the characteristics of the data, the value used in this case to show centrality will be the **median**. The assumption is that the most frequent value for easy words is lower than for difficult words.

[11] A boxplot is a graphical representation of data that takes into account its quartiles.

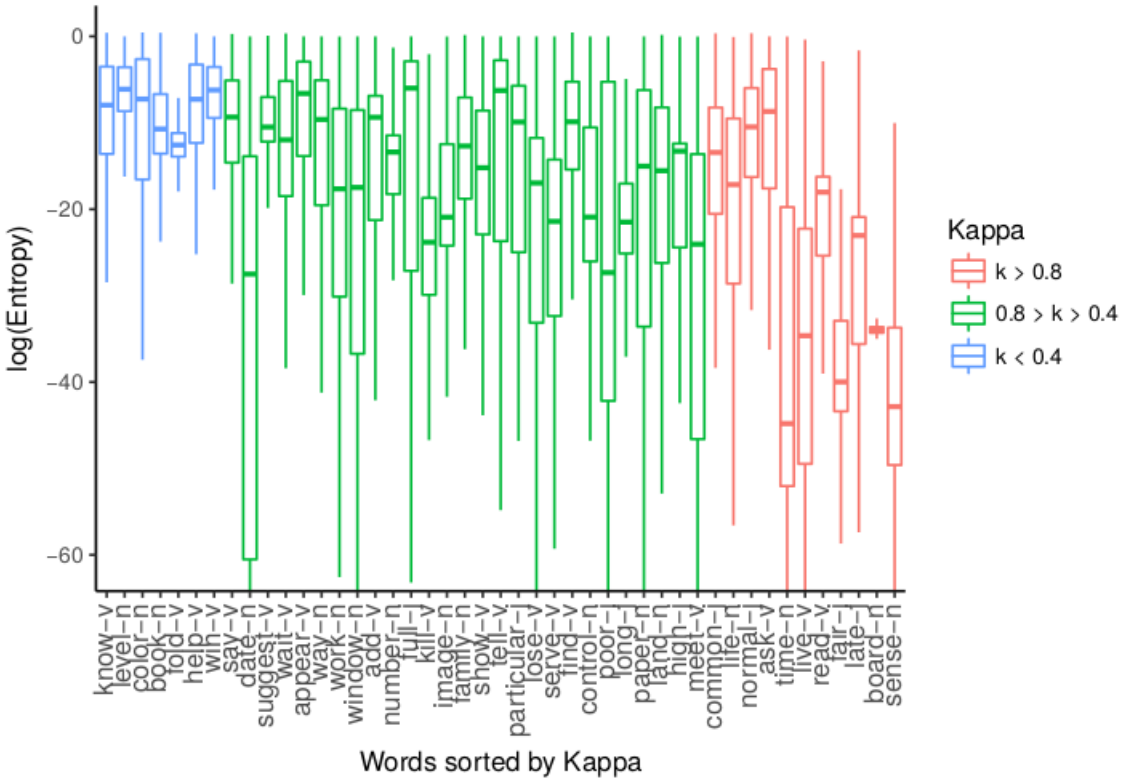


Figure 9: Boxplots of the annotation entropies, for each word in the dataset. Words are sorted by their kappa value, in ascending order.

Figure 9 shows that the words with a **higher kappa value** (in red) have their interquartile range (IQR) in **lower ranges of entropy**, which means that are **overall less difficult to annotate**.

Regarding the words with a **lower kappa value** (in blue), the interquartile ranges are tighter than in words with a higher kappa value, and these ranges lie on the **higher values of entropy**, meaning that they include a **high number of difficult annotation examples** and are, thus, **more difficult to annotate** in general.

These observations are more evident in the most extreme words in the dataset: ***sense-n*** and ***know-v***, being the word with the **highest kappa value** (0.96) and the word with the **lowest kappa value** (0.13), respectively, with the following considerations regarding their distributional characteristics:

- Regarding **entropy ranges**, 90% of the log-entropy for ***sense-n*** ranges between $[-66.3, -15.4]$, whereas the range for ***know-v*** is considerably higher, with values between $[-26.9, 0.5]$, reinforcing the observations in Figure 9.

- Regarding the **most frequent value**, represented by the **median**, the values are around -42.8 for ***sense-n*** and around -7.9 for ***know-v***.

In order to visualize the considerations made above, Figure 10 overlaps the histograms for ***sense-n*** and ***know-v***:

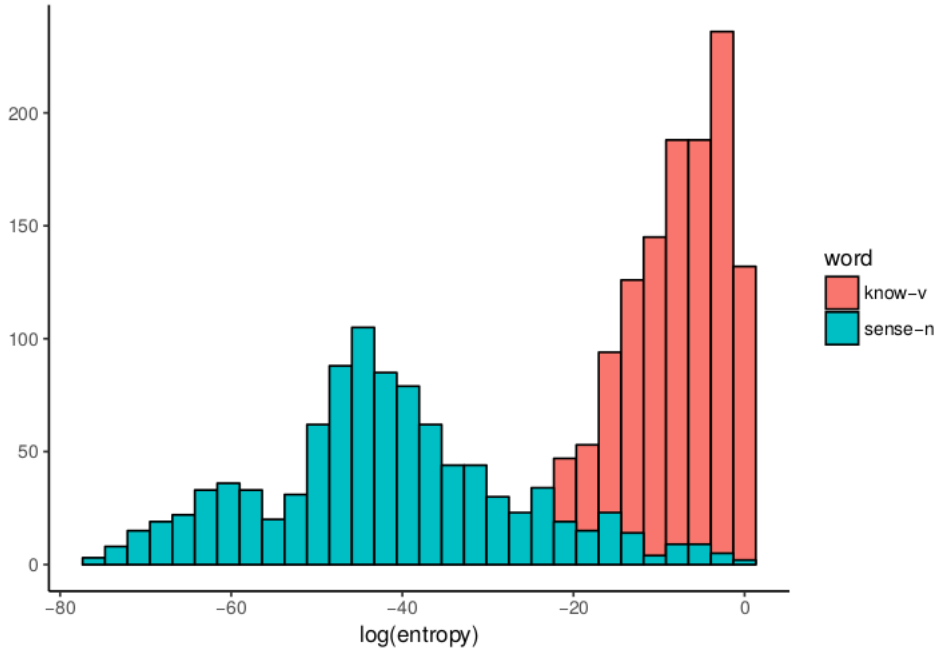


Figure 10: Overlapped histograms of the word with the highest kappa value (***sense-n***), in blue, and the word with the lowest kappa value (***know-v***), in red.

Figure 10 shows in a very intuitive way the distribution of both words. It can be observed that the distribution for ***know-v*** is very skewed, whereas the distribution for ***sense-n*** is much less skewed.

By taking into account the observations showed above, it can be considered that there exists a correlation between the difficulty of target words (measured with kappa agreement) and the difficulty of examples (measured with entropy). More specifically, examples for *easy* target words tend to lie in lower ranges of entropy, with lower medians, and examples for *difficult* target words tend to lie in higher ranges of entropy, with higher medians.

3.4.3 Conclusions

In this section, the supervised measures to model difficulty of target words (kappa agreement) and examples/context sentences (entropy) have been analysed together, in order to determine if there is any relationship between them and validate them as supervised difficulty metrics for target words and examples/context sentences. The assumption for doing so is that the difficulty of a target word is associated to its context sentences, and,

therefore, an *easy* target word should have a low overall entropy, whereas *difficult* target words should have a high overall entropy.

Each target word's entropy distribution for its examples has been analysed in terms of two characteristics: **entropy range**, extracted from the boxplots in Figure 9, created for each word, and the **most frequent value**, extracted from the median of the distribution.

The analysis of the target word with the highest kappa value (***sense-n***) and the target word with the lowest kappa value (***know-v***) have proved the assumptions, since ***sense-n*** shows a lower entropy range and median than ***know-v***.

In conclusion, it can be concluded that this initial peek at the data reinforces the assumption that it is possible to characterize difficulty of target words and examples by using kappa agreement and entropy, respectively.

4 Estimation of difficulty without using annotated data

As it has been seen in the previous section, it is possible to determine the difficulty of a target word or an example by taking into account the annotation data from a corpus. However, the main objective of this project is to **predict** this difficulty without making use of the information that comes from the annotation task.

In Section 1.2, it has been observed that sense definitions for some target words present **similarities** between them, making the disambiguation task difficult. Furthermore, some context sentences seemed to be **underspecified**, in the sense that they did not provide enough information for a target word to be disambiguated.

In this section, and following the observations in Section 1.2, two unsupervised difficulty measures will be proposed: **similarity between sense definitions** for **target words** and **probability** (as a metric to measure specificity), for **context sentences**. In order to prove their validity as difficulty measures, both supervised and unsupervised metrics will be **correlated**.

4.1 Calculating relations between variables: correlation

As it has been proved in Section 3, kappa and entropy are valid supervised metrics to model difficulty of target words and context sentences, respectively. In this section, these two measures will be **correlated** to the proposed unsupervised measures. Thus, it is necessary to describe the correlation methods that will be used in this thesis.

Before going into correlation measures, some important concepts need to be explained:

- **Covariance**. Covariance is a measure that quantifies how two variables change together.
  The covariance value can be calculated with the following formula:

$$cov(X,Y) = \frac{\sum E[(X - E[X])(Y - E[Y])]}{n-1} \tag{21}$$

  Where:

  - X and Y are variables
  - E[X] and E[Y] are the expected values of X and Y, respectively, that correspond to the mean of all the values in X and Y, respectively.
  - n is the number of items in the dataset

  According to the covariance value, there are three different scenarios: **negative** covariance, **no** covariance and **positive** covariance. The closer the value to 0, the less covariance.
  The difference between the different covariance scenarios can be summarized in the following image:

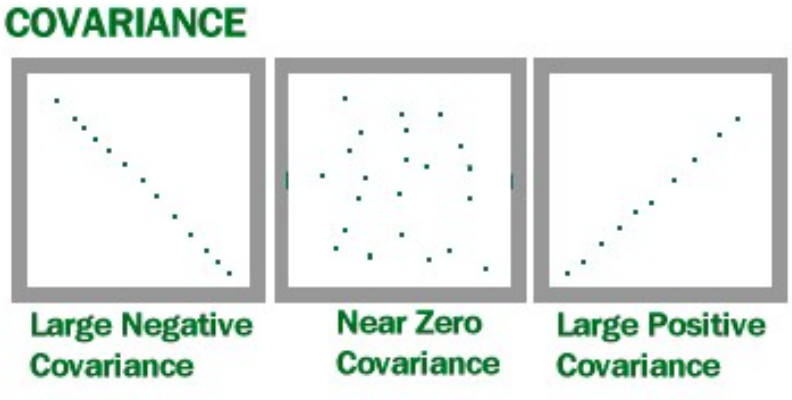


Figure 11: Covariance scenarios (Glen 2013)

Thus:

- **Negative covariance**: the values of Y variable decrease when the values of X increase.
- **No covariance**: there is no pattern.
- **Positive covariance**: the values of Y increase when the values of X variable increase.

However, the values that can be returned from this measure are not limited to any set of values (for example, -1 to 1), so it can be difficult to extract conclusions by having a covariance value. So as to be able to determine if the value is low or high, the **standard deviations** for each set of variables can be used.

- **Standard deviation**. It is a measure that reflect up to what extent the variables obtained are spread out close (or not) to the average. In order to obtain the standard deviation, three steps are required:

  1. Get the **mean** from all the values:

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{22}$$

  2. Get the **variance**:
     - Measure the *distance* from the mean for each variable:

$$distance = x - \bar{x} \tag{23}$$

     - Square the *distance* for each value of the variable:

$$squaredDist = (distance)^2 \tag{24}$$

– Average the *squaredDistance* with the rest of values:

$$avg = \frac{squaredDist_1 + squaredDist_2 + \cdots + squaredDist_n}{n} \tag{25}$$

3. Get the squared root of the average:

$$sqrtAvg = \sqrt{avg} \tag{26}$$

To wrap up, standard deviation can be summarized in the following formula:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{27}$$

- **Monotonicity**. Monotonicity stands for a relationship between two variables that behave according to one of the following patterns:
  - When X increases, Y decreases
  - When X increases, Y increases

  Also, if the tendency is monotonic, it can also be **linear** (that is, the relation between the variables can be shaped as a line).
  The different scenarios can be seen in the following image:

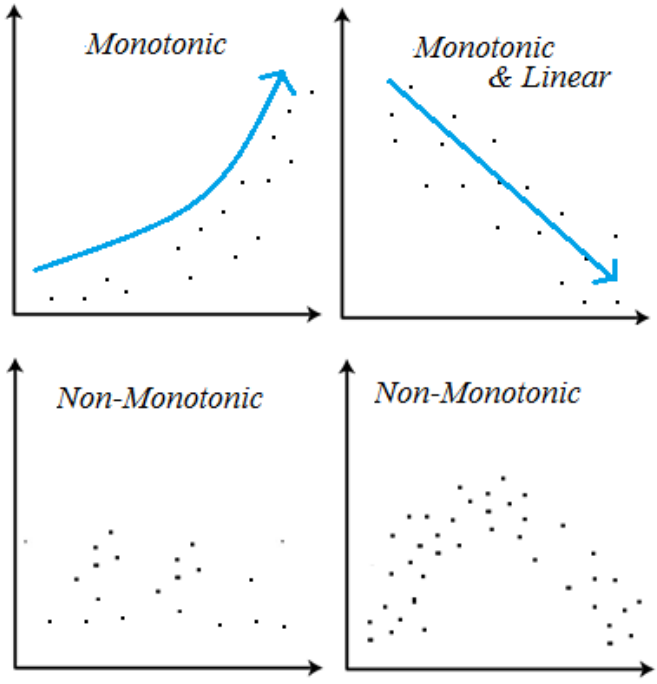


Figure 12: Monotonicity scenarios (Glen 2017)

Note that monotonicity is not the same as covariance: monotonicity only takes into account if there is a tendency between the variables (without making any distinctions on its direction), whereas covariance does make distinctions on the direction. Monotonicity can be calculated by a product-moment correlation such as **Pearson's correlation coefficient**. Additionally, if the data is monotonic and linear[12], the best coefficient is **Spearman's correlation coefficient**.

There are different measures that provide a correlation coefficient between variables, distributed in two categories: **product moment** coefficient and **rank** coefficients:

- A **product moment** coefficient takes into account the **covariance** and the **standard deviation**. This coefficient determines "the strength and direction of the **linear** relationship between two variables"[13]. In other words, it tries to answer the question *"Can I draw a line to represent the relationship between the two variables?"* The product moment coefficient is also known as **Pearson product-moment correlation coefficient**, or **Pearson correlation coefficient**.

- **Rank** coefficients do not take into account the parameters in relation to the totality of the dataset (that is, standard deviation or mean). For this reason, they are considered as **non-parametric** measures.
  Rank coefficients, such as **Spearman's correlation coefficient** or **Kendall's correlation coefficient**, do not measure the correlation by taking into account the linear relationship between the variables, but the **monotonicity**.

4.2 Estimating target word difficulty

The **second hypothesis** states that it is possible to determine the difficulty of a target word by analysing the **similarity** between all the pairs of senses of said word. In fact, it is assumed that the higher the similarity between the senses of a target word, the more difficult to annotate. These assumptions come from the simple observation of the sense definitions of the words in the dataset:

WordPos	SenseId	Definition
add-v	1	make an addition (to); join or combine or unite with others; increase the quality, quantity, size or scope of
add-v	2	state or say further
add-v	3	bestow a quality on
add-v	4	make an addition by combining numbers
add-v	5	determine the sum of
add-v	6	constitute an addition

Table 6: Senses for ***add-v***

[12] In order to determine if the data is monotonic and/or linear, a scatter plot with all the data in the dataset must be used.

[13] *Spearman's Rank-Order Correlation* (n.d.)

As it can be seen in Table 6, senses 1, 4, 5 and 6 have a similar meaning. Hence, in order to prove the **second hypothesis** of this study, the senses of the words in the dataset will be analysed by calculating the **similarity** between the combinations of their sense definitions, and obtaining an average similarity value for each word. Also, the results obtained will be correlated to their kappa agreement, to determine if the similarity between sense definitions is a determining factor to estimate the difficulty of disambiguation of target words.

4.2.1 Calculating similarity

In order to calculate similarity between a pair of sense definitions, two algorithms will be used: one based on **overlap** methods (bag-of-words with cosine similarity) and another based on **embeddings** (GloVe word embedding-based centroid algorithm (Pennington, Socher, and C. D. Manning 2014) with cosine similarity).

The reason for choosing both algorithms is simple: the overlap-based algorithm is a simple approximation to retrieve the similarity between pairs of words/sentences, whereas GloVe, as a more complex algorithm, is a modern solution which has proven to have better results (Pennington, Socher, and C. D. Manning 2014) than other similarity algorithms.

- **Overlap. Bag-of-words with cosine similarity**
  This approach makes use of two metrics in order to obtain a similarity value between the elements of a pair (in this case, sense definitions): the bag-of-words model and cosine similarity.
  The bag-of-words model assumes that each document (in this case, each sense definition), is a bag of words. That is, the order is not significant, only the words and their presence in the text. With those retrieved words, features are extracted, being the most common the term frequency or term presence.
  In order to compare the similarity between a pair of sense definitions, a vector for each word in the sense definitions will be created, by taking into account the term presence in each of them. The vector contains the information stored in a binary format: 1 for "present", and 0 for "absent". For example, given these two sentences:

  (1) John likes to watch movies. Mary likes movies too.
  (2) John also likes to watch football games.

  The list of words would be as follows:

  ["*John*", "*likes*", "*to*", "*watch*", "*movies*", "*Mary*", "*too*", "*also*", "*football*", "*games*"]

  The vectors generated by taking into account the term frequency would be the following:

  (1) [1, 1, 1, 1, 1, 1, 1, 0, 0, 0]
  (2) [1, 1, 1, 1, 0, 0, 0, 1, 1, 1]

To establish how similar are the sense definitions, the cosine of the angle between the vectors will be calculated. The cosine similarity can be summarised in the following formula:

$$cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} \tag{28}$$

In a nutshell, the cosine similarity calculates the cosine of the angle between the vectors. Depending on the angle, the function will return a value between -1 and 1[14]. Each value states for the different cases of similarity:

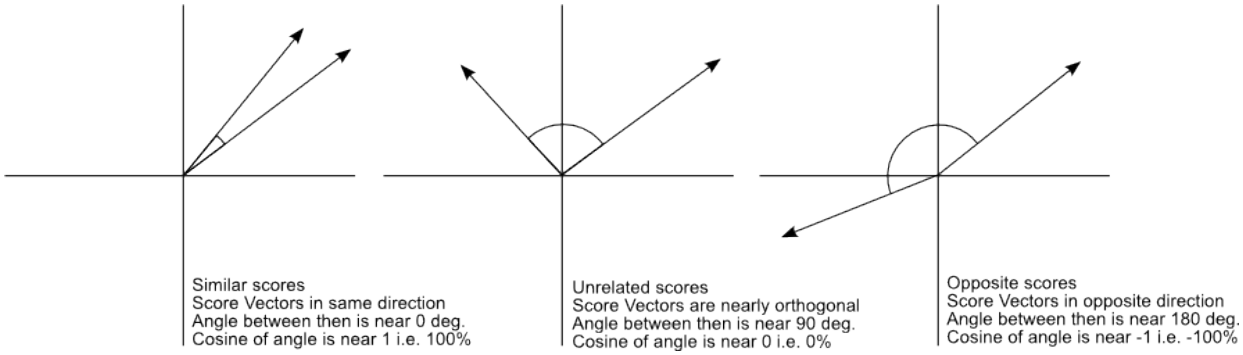


Figure 13: Cosine similarity value scenarios (Perrone 2013)

- **Embeddings. GloVe word embedding-based centroid algorithm with cosine similarity**
  As in the previous section, this approach also makes use of two metrics: the Stanford's GloVe word embedding-based centroid algorithm and, like in the previous approach, cosine similarity.
  GloVe (from Global Vectors), according to Pennington, Socher, and C. D. Manning (2014), "is an unsupervised learning algorithm for obtaining vector representations for words". The training, according to the authors, "is performed on aggregated global word-word co-occurrence statistics from a corpus". That is, all the words in the corpus are stored in terms of co-occurrence to create a global vector space.
  For each sense definition, a vector is obtained taking into account the embeddings of each word and storing the centroid of those embeddings. After the vector for each sense definition is obtained from the centroid of the embeddings for its words, the similarity of the two vectors is calculated by using the cosine function.

4.2.2 Experiment design

In order to prove the hypothesis, an experiment, in which the similarity between the sense definitions of the words will be analysed, will be performed.

[14]In this project, the values for similarity have been scaled to return values between 0 and 5.

The experiment will consist of the following steps:

- All the possible combinations of sense definitions for a word will be created, in order to be able to create the input for the scripts that calculate similarity.

- After having the input file(s)[15] generated, they will be feed to two scripts that will calculate the similarity between each pair. The first will calculate the similarity by using an **overlap algorithm**, while the second will calculate the similarity by using an **embedding-based algorithm**. Both algorithms have been explained in the previous section.
  Each algorithm will return a value between 0 and 5, being 0 the value for no similarity and 5 the value for total similarity. Table 8 is an example.

- After having generated all the similarities between the combinations of sense definitions, the average similarity for each word (by each algorithm) will be calculated. This procedure helps establishing a single measure for each word, in order to simplify the analysis.
  The output has the structure shown in Table 7.

word	**overlap_avg_sim**	**embeddings_avg_sim**	**Kappa**	**Number of senses**
find-v	0.571	3.851	0.73	16
date-n	1.044	3.852	0.49	8
wait-v	0.167	3.578	0.56	4
fold-v	0.372	3.762	0.37	5
ask-v	0.849	4.21	0.83	7

Table 7: Output sample of average similarities for each algorithm, along with kappa value and number of senses for each word.

- So as to be able to prove the initial hypotheses, the values generated for each word by each algorithm will be correlated to the results obtained by estimating the difficulty of words with annotation data. In other words, the similarity values of each word, for each similarity algorithm, will be correlated to the corresponding kappa value extracted from the annotations.
  To do so, two correlation algorithms, mentioned in Section 4.1, will be used: **Pearson** and **Spearman**, whose results will be analysed in the next section in order to extract conclusions.

[15]Since the requirements for each script are the same, the input for both will also be the same.

		Overlap similarity	Embeddings similarity
make an addition to join or combine or unite with others increase the quality quantity size or scope of	state or say further	1.5	4.04
make an addition to join or combine or unite with others increase the quality quantity size or scope of	bestow a quality on	0.5	3.87
make an addition to join or combine or unite with others increase the quality quantity size or scope of	make an addition by combining numbers	1.22	4.32
make an addition to join or combine or unite with others increase the quality quantity size or scope of	determine the sum of	1.0	4.01
make an addition to join or combine or unite with others increase the quality quantity size or scope of	constitute an addition	1.15	3.88
state or say further	bestow a quality on	0.0	3.07
state or say further	make an addition by combining numbers	0.0	3.63
state or say further	determine the sum of	0.0	3.55
state or say further	constitute an addition	0.0	3.43
bestow a quality on	make an addition by combining numbers	0.0	3.33

Table 8: Fragment of input file (for ***add-v***) and associated similarity values

4.2.3 Results

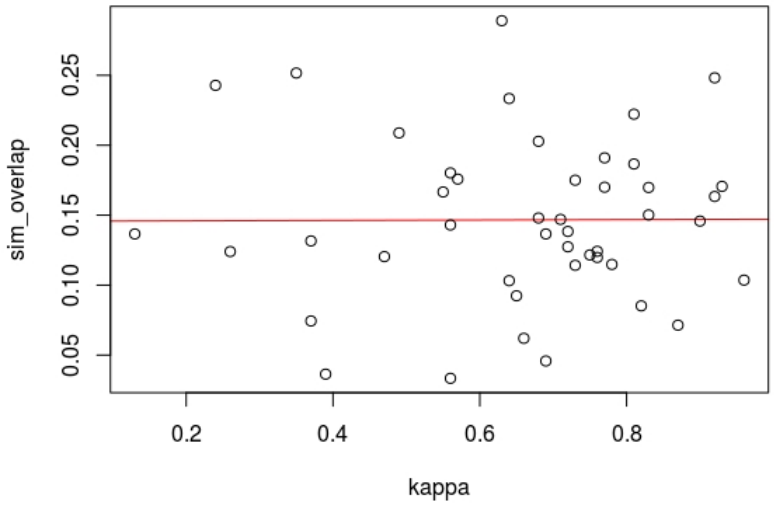


Figure 14: Correlation plot - **kappa agreement** and **overlap similarity**. **Pearson** = 0.00513121, **Spearman** = 0.02395861.

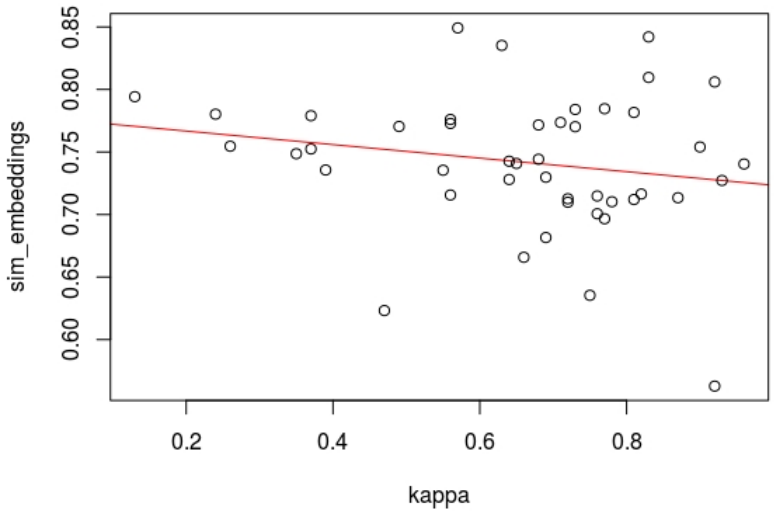


Figure 15: Correlation plot - **kappa agreement** and **embeddings-based similarity**. **Pearson** = -0.192309, **Spearman** = -0.1852694.

Figure 14 shows a scatter plot, in which **kappa** (in the X axis) and **overlap similarity** (in the Y axis) are correlated. In order to analyse the correlation between the two metrics and observe the monotonicity and linearity of the data, a linear regression (marked in red) has been added.

Similarly, Figure 15 also shows a correlation plot, in which **kappa** (in the X axis) is correlated with **embeddings similarity** (in the Y axis). Also, the linear regression line has been marked.

As it can be seen in Figure 14, the correlation between the kappa agreement and the overlap similarity is very close to 0. Therefore, if the overlap similarity value is taken into account, the two measures (kappa and overlap similarity) have no relationship at all. This result may be due to the fact that several pairs of sense definitions obtained a 0 similarity value −because of the simplicity of the algorithm−, and the results may be distorted, since it may decrease the values for similarity when averaging.

On the other hand, Figure 15, in which **embeddings-based similarity** is analysed, shows better results. In this case, the correlation is negative, so **the higher the agreement, the lower the similarity**.

The results from this second plot show signals that similarity between senses is a valid measure to determine the difficulty of a target word. More specifically, a **target word** is **easy** when the average **similarity** is **low**.

However, the value of the correlation measures is not high enough to determine that the two measures are highly correlated.

4.2.4 Conclusions

After assuming in the previous section that **kappa** agreement can be used to estimate the difficulty of target words using annotation data, the main purpose of this section was to estimate the difficulty of target words **without using annotation data**.

The starting point was the **second** hypothesis, in which it was claimed that **similarity**, extracted from the **sense definitions** of the word is a valid measure to **predict** the difficulty of a target word.

In order to prove the hypothesis, the similarity between sense definitions of each word has been calculated by using two different algorithms: **overlap-based**, as a simple approximation, and **embeddings-based**, as a more complex and modern solution. After calculating the similarity of all the pairs of sense definitions for each word, a mean value has been extracted, in order to provide each word with a single value of sense definition similarity.

So as to prove the initial hypothesis, the kappa agreement values for each word have been correlated to their corresponding similarity values, for each similarity algorithm. To do so, two correlation algorithms have been used, **Pearson's correlation coefficient** and **Spearman's correlation coefficient**.

The final results have been different for each similarity algorithm. On the one hand, the

correlation between kappa and overlap similarity has been very close to 0, that is, showed no relation between both values. Although it may seem a bad result, it was somehow expected, since it is a simple algorithm and showed a similarity of 0 in many pairs of sense definitions, which can distort the results significantly.

On the other hand, the correlation between kappa and embeddings similarity has showed positive results, since it has provided a negative correlation between both values, meaning that the higher the agreement, the lower the similarity. Thus, a word is **easy** when the average **similarity** value for said word is **low**.

However, although the results have been positive for this algorithm, it is true that the values of correlation are not high enough to state that there is a high correlation.

Furthermore, averaging the similarities for each sense definition may mask problematic cases. For example, a word with several senses and only two similar sense definitions may show a low similarity value, caused by the low similarity between the other sense definitions. In order to solve this problem, a new generalization may be proposed in the future.

Nevertheless, the results obtained show signals that definition similarity is a factor that is related to the difficulty of words.

4.3 Estimating context sentence difficulty

According to Koirala and Jee (2015), the difficulty when understanding sentences mainly resides in two features: **length** and the **frequency** of the words in the sentence. In this case, the goal is not to model the difficulty of understanding context sentences, but to discern if the context sentence contains enough information in order to be disambiguated, but it is possible to apply these same factors to the analysis performed in this project.

In this section, thus, all the context sentences involved in the annotations will be analysed in terms of their length and their overall frequency, as an approximation to the results obtained by the authors, adapted to the necessities of this thesis.

The assumption is that **probable** contexts, since they include **more frequent** (and, thus, **less specific**) words, will be **difficult** to annotate, since the context does not provide enough information, i.e., it is **underspecified**.

Similarly, **unprobable** contexts, since they include **less frequent** (and, thus, **more specific** words), will be **easy** to annotate.

It is expected, thus, that there is some correlation between the entropy of an example (i.e., its difficulty) and the probability of the context. In this sense, the **higher the entropy** (that is, the less agreement between annotators in the annotations), the **higher the probability** of the context, being probability a **useful factor** to determine the difficulty of a sentence or, at least, one of the factors involved in it.

4.3.1 Calculating probability of sentences

In order to calculate the probability of a sentence (that is, its probability of occurrence), a **language model** will be used. A **language model** is a model whose function is to calculate the probability of a sequence of words. In other words, it estimates the probability of occurrence of a given sequence of words in a given language. There are two types of language models:

- **Probabilistic language models**. As their name suggests, probabilistic language models make use of statistics in order to calculate the probability of a sequence of words. These models can be of two types:

  - ***Unigram* models**. These models are based on the probability of each of the words that compose the sentence in the corpus. The probability of a sequence of words is the product of the probability of each of the words in the sentence:

    $$P_{\text{uni}}(t_1 t_2 t_3 t_n) = P(t_1)P(t_2)P(t_3)\cdots P(t_n) \tag{29}$$

    For example, given the sentence "The black cat eats apples":

    $$\begin{gathered} P(\text{The}) = 0.647 \\ P(\text{black}) = 0.212 \\ P(\text{cat}) = 0.354 \\ P(\text{eats}) = 0.435 \\ P(\text{apples}) = 0.239 \end{gathered}$$

    To calculate the probability:

    $$P_{\text{uni}}(the, black, cat, eats, apples) = P(the)P(black)P(cat)P(eats)P(apples) \tag{30}$$

    Thus:

    $$P_{\text{uni}} = 0,005 \tag{31}$$

    The main problem of these models in terms of probability of sentences is that the probability of the combination of said words is not contemplated, only the probability of appearance of each word. However, these models are useful in other systems, such as spell correction (for example, the word *minutes* would have a higher probability than *minuets*, a word that does not exist in English and, therefore, will have a low probability).

  - ***n-gram* models**. *n-gram* models base the probabilities of a word in the previous n-1 words. There are many types of *n-grams*: **bigrams** (the probability of a word depends on the previous word), **trigrams** (the probability depends on

the two previous words), and so on.
For example:

* **Bigram**

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-1}) \tag{32}$$

For example, given the sentence ("The man walks home")[16]:

· $P(\text{The man walks home}) =$

$$\begin{gathered} P(\text{The} \mid *) \cdot \\ P(\text{man} \mid * \text{ The}) \cdot \\ P(\text{walks} \mid \text{man}) \cdot \\ P(\text{home} \mid \text{walks}) \cdot \\ P(\text{home} \mid \text{STOP}) \end{gathered}$$

* **Trigram**

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_{i-2}, w_{i-1}) \tag{33}$$

For example, given the same previous sentence ("The man walks home") [17]:

· $P(\text{The man walks home}) =$

$$\begin{gathered} P(\text{The} \mid * *) \cdot \\ P(\text{man} \mid \text{The}) \cdot \\ P(\text{walks} \mid \text{The man}) \cdot \\ P(\text{home} \mid \text{man walks}) \cdot \\ P(\text{home} \mid \text{walks STOP}) \end{gathered}$$

- **Neural language models**. Another approach in order to estimate the probability of a sequence is to use neural networks. Although they are more complex than probabilistic models, these models achieve better results.

[16]In order for the bigram model to be successful, it is considered that there is one unit before the first word of the sequence. It is also considered that after the last word of a sequence there is an extra unit, represented as STOP.

[17]In order for the trigram model to be successful, it is considered that there are two units before the first word of the sequence. As in the previous case, it is considered that after the last word of a sequence there is an extra unit, represented as STOP.

In this project, a probabilistic model, **KenLM**[18], will be used. This model has been trained by using the corpora from the 8th Workshop on Statistical Machine Translation (WMT, 2013)[19]. In it, 84 million segments and 2,000 million words are included, extracted from bilingual corpora English-Spanish (the part corresponding to English) and monolingual English corpora. These corpora include proceedings from the European Parliament, news, and Internet-extracted data from various domains.

4.3.2 Experiment design

The experiment performed in order to prove the hypothesis will consist of the following steps:

- First of all, a set of preprocessing steps will be performed:
  - As a first preprocessing step, the sentences will be **tokenized**. **Tokenization** is a very important process in most NLP tasks, in which a sentence is split in its sub-elements. Loosely speaking, tokens are often considered as the words that form the sentence, although a token strictly corresponds to an "instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing"[20].
    For example, the previous sentence:

    "*and they'd have to take several projects to go up there before and add on piece by piece*"

    Would become:

    "*and* | *they* | *'d* | *have* | *to* | *take* | *several* | *projects* | *to* | *go* | *up* | *there* | *before* | *and* | *add* | *on* | *piece* | *by* | *piece*"
  - Finally, the tokens undergo a process of **truecasing**. Truecasing is a process in which all capital letters are removed from the sentences, except for the proper nouns.
    For example, given the tokenized sentence:

    "She | *is* | *Maria* | *'s* | *sister"*

    Would become:

    "she | *is* | *Maria* | *'s* | *sister"*
- After preprocessing the sentences, they will be fed to the language model, which will return the probability for each of the context sentences.

[18] https://kheafield.com/code/kenlm/
[19] http://www.statmt.org/wmt13/translation-task.html
[20] C. D Manning, Raghavan, and Schütze (n.d.)

- Also, the length for each sentence will be extracted from its number of tokens.
- In order to prove the hypotheses, the probabilities for each context sentence will be correlated to their corresponding entropy value.

Regarding the last step, due to the large amount of data, the data is **noisy**. However, as Yarowsky and Florian (2002) work shows, the fact that a considerable number of annotators is involved suggests that the estimations will be more robust.

Nevertheless, it is possible to reduce the noisiness of the annotation data by grouping the data according to its length and averaging it. This approach comes from the fact that probability and length are **highly** correlated, as the following plot shows:

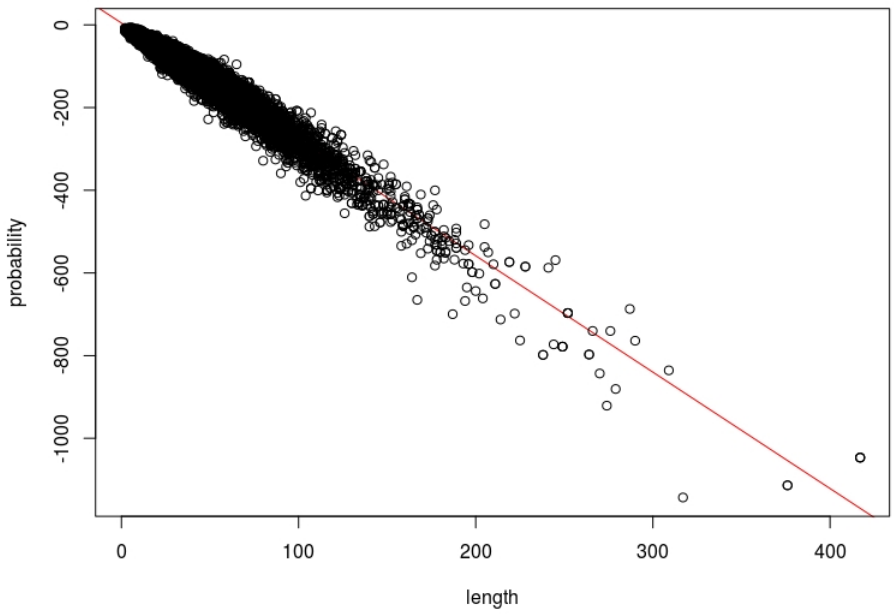


Figure 16: Correlation plot - **probability of occurrence** and **length of context sentence**

In this case, the entropy values and the probabilities of each example will be grouped according to their length, and then the values of probability/entropy for each length will be averaged, in order to reduce the noise that other variables that are not being taken into account may cause.

It is expected to obtain a representation that suggests that the correlation does exist and, thus, prove that probability may help predicting the difficulty of an example without annotation data.

Furthermore, and since it has been pointed in Section 3.3.3, it is possible to model the

difficulty of context sentences by using entropy, and the relationship between entropy and difficulty is more evident when the words are easy. For this reason, the experiment will be performed only taking into account ***easy* words**, in order to make sure that the results obtained with all the set of words are accurate enough and to reinforce the hypothesis.

For that matter, those words with a kappa value **equal or higher than 0.8** will be considered as ***easy***.

4.3.3 Results

In Figure 17, in order to make the correlation analysis easier, a logarithmic transformation has been applied to the variables[21]:

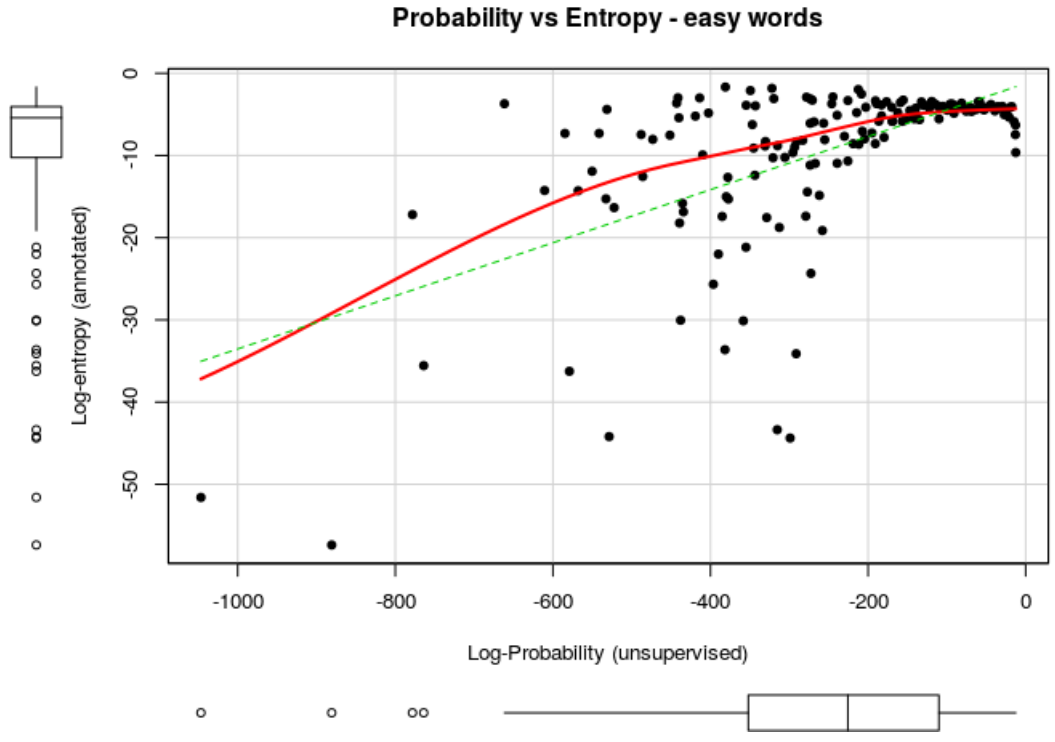


Figure 17: Correlation plot - **entropy** and **probability of occurrence** of ***easy*** words, grouping data - rescaled. **Pearson** = 0.6016828, **Spearman** = 0.4949125

Figure 17 shows a correlation plot between log-probability (in the X axis) and log-entropy (in the Y axis). Each point represents a set of context sentences, which have been grouped according to their length and having their entropy values averaged. Furthermore, a linear regression line has been added in order to be able to observe the linearity of the data and its behaviour in a clearer way. Also, the values for Pearson and Spearman are indicated.

[21] In fact, the transformation for probability is applied and returned by the language model.

4.3.4 Analysis and conclusions

The main purpose of this section is to prove that it is possible to determine the difficulty of a sentence without annotation by taking into account two factors: its **length** and its **probability**. The first has been computed as the length of the list of tokens from the sentence, whereas the second has been computed using a language model.

An experiment has been carried out in order to be able to determine the correlation between **entropy** (a supervised factor for determining difficulty) and **probability** (the proposed unsupervised factor for determining difficulty of context sentences). Taking into account that the annotation data is noisy, due to the huge amount of data, the data has been grouped according to length (strongly correlated to probability) in order to reduce noise.

Due to the skewness of the data, it has also been rescaled in order to better visualize and analyse the resulting plots.

The results in Figure 17 show that there is some evidence that probability is correlated to entropy. More specifically, it seems that the **higher the entropy**, the **higher the probability**, a behaviour that **supports** the hypothesis, which is also confirmed by taking into account the values for **Pearson** and **Spearman**, of **0.5** and **0.6**, respectively, which show that the correlation is apparently **moderate**, although it is important to take into account that grouping the data and scaling the entropy value has caused the values for Pearson and Spearman to increase. Therefore, it will be considered that there are **signals** of the relationship between probability and entropy (and, therefore, the difficulty) of the context sentence.

Thus, and taking into account these considerations, the plots have shown the **expected behaviour**, since they have proven that probability can be a factor that can **help** modelling difficulty in context sentences without relying on annotations, although it is not the only factor to be taken into account, considering that the interaction between the two measures has not been strong enough.

In conclusion, the results of the first experiment have **supported** the ideas stated in the hypothesis, in which it has been considered that the probability of a sentence is a valid factor to determine its difficulty. However, as the values for Pearson and Spearman have shown, the results of the first experiment show a **moderate correlation** between the supervised difficulty factor (entropy) and the proposed unsupervised difficulty factor (probability).

Furthermore, it has been proved that probability is a factor that **helps** modelling the difficulty of a context sentence, but not the only one to be taken into account.

5 Predicting example difficulty

After having determined that the proposed predictors −similarity between sense definitions and probability of context sentences− help model the difficulty of words and context sentences, respectively, it is necessary to analyse the interaction of both measures together with the overall difficulty of the example (that is, the word in context), which is modelled supervisedly by the **entropy** of the annotation task.

5.1 Experiment design

The purpose of this experiment is to design a model capable of predicting, by using the proposed unsupervised measures (similarity for words, and probability for context sentences), the difficulty of an example. That is, the goal is to be able to predict the difficulty of an example by combining the proposed predictors in a model capable of doing so.

In order to be able to design the model, a **linear regression** approach will be used, since it is a widely used approach in order to make predictions between an **explicative** variable (the value to be predicted) and one or more **predictors** (the variables used to predict the explicative variable). Depending on the characteristics of the data, the predictors can be centered to their mean, meaning that the data will be fitted in order to have 0 as their mean, in order to better interpret the results.

In a linear regression approach with multiple variables, the goal is to solve the following equation:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_k X_{ik} + \epsilon_i, \text{ for } i = 1, \ldots, n \tag{34}$$

Where:

- y_i = the **explicative variable**.
- β_0 = the **intercept**, which is the predicted value for y_i when the values for the predictors are 0. If the predictors have been centered, the value for 0 is, in fact, the average value for each of the predictors.
- X_{ik} = the **predictor** k from y_i.
- β_k = the **slope** of X_{ik}. The **slopes** for each predictor will be the increasing of y_i when the value for the predictor increases by 1. For example, if the estimated slope for $X_i k$ is 0.03, it means that when the value for $X_i k$ increases by 1, the y_i increases by 0.03.
- ϵ_i = the **error** of the estimation for the annotation.

In this equation, the unknowns would be the **intercept** and the **slopes**, which will be estimated by the model.

In this approach, three variables will be combined: **entropy** (as the supervised measure to establish example difficulty), as the **explicative** variable, and **averaged sense similarity**[22] −calculated with **word embeddings**− [23] and **probability of the context sentence** −obtained from a **language model**− (as the unsupervised measures to predict the difficulty of words and context sentences), as the **predictors**, in order to quantify their relationship by using the coefficients obtained. Also, the interaction between the predictors will be taken into account.

The proposed model will be defined by having the **predictors centered** and the **entropy log-scaled**, in order to better interpret the results.

Therefore, the equation to be resolved is the following:

$$Entropy = Intercept + Slope_{Sim} * Sim + Slope_{Prob} * Prob + Slope_{SimProb} * Sim * Prob + Error \quad (35)$$

After obtaining the coefficients, a plot will be created in order to visualize the results and be able to extract conclusions.

5.2 Results and analysis

In order to be able to perform the analysis, it is necessary to create the necessary variables to do so. For this reason, the predictors values will be centered to their mean, obtaining two new variables, `c.probability` and `c.embeddings_sim`, for probability and similarity, respectively. Furthermore, the entropy will be log-scaled, obtaining the variable `log.entropy`.

Figure 18 shows important descriptive data. For each of the variables, the quartile values are shown, along with the minimum and maximum values (that is, the **range**) and the mean. It is worth comparing the *raw* values and their centered equivalents, where the means are 0. For entropy, it can be observed that the log-scaled values are more suitable for analysis.

In this case, it is important to take into account the mean for each of them, in order to be able to extract conclusions.
After log-scaling entropy and centering the predictors, the linear regression is calculated.

[22] *vide* Section 4.2
[23] *vide* Section 4.3

```
##       entropy           probability     embeddings_sim
##  Min.   :0.0000000   Min.   :-1143.17   Min.   :2.814
##  1st Qu.:0.0000000   1st Qu.: -107.57   1st Qu.:3.567
##  Median :0.0000005   Median :  -76.48   Median :3.713
##  Mean   :0.0377820   Mean   :  -88.70   Mean   :3.723
##  3rd Qu.:0.0006400   3rd Qu.:  -52.09   3rd Qu.:3.895
##  Max.   :1.5707386   Max.   :   -6.12   Max.   :4.246
##     log.entropy        c.probability    c.embeddings_sim
##  Min.   :-175.4843   Min.   :-1054.47   Min.   :-0.90881
##  1st Qu.: -27.6962   1st Qu.:  -18.88   1st Qu.:-0.15581
##  Median : -14.5458   Median :   12.21   Median :-0.00981
##  Mean   : -20.2881   Mean   :    0.00   Mean   : 0.00000
##  3rd Qu.:  -7.3541   3rd Qu.:   36.61   3rd Qu.: 0.17219
##  Max.   :   0.4516   Max.   :   82.58   Max.   : 0.52319
```

Figure 18: Statistics for entropy (explicative variable) and similarity and probability (predictors), raw and scaled

```
## Call:
## lm(formula = log.entropy ~ c.probability + c.embeddings_sim +
##     c.probability:c.embeddings_sim, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -154.610   -7.298    5.648   12.714   25.382
##
## Coefficients:
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                    -20.282578   0.086369 -234.836  < 2e-16 ***
## c.probability                   -0.006465   0.001413   -4.577 4.73e-06 ***
## c.embeddings_sim                 4.588766   0.325757   14.086  < 2e-16 ***
## c.probability:c.embeddings_sim   0.033354   0.006328    5.271 1.36e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.26 on 44676 degrees of freedom
## Multiple R-squared:  0.006059,   Adjusted R-squared:  0.005992
## F-statistic: 90.78 on 3 and 44676 DF,  p-value: < 2.2e-16
```

Figure 19: Statistics for the linear regression and the interactions between the explicative variable (entropy) and the predictors (similarity and probability)

Figure 19 shows the results obtained by the linear regression, as a result of the sum of the predictors and the interaction between them.
For the intercept, each predictor, and the interaction between predictors, four coefficients are obtained:

- **Estimate**. It indicates, when the value of the predictor increases by 1, how much the value for entropy increases. For example: if the estimate for probability is 0.03, it means that when probability increases by 1, the value for entropy increases by 0.03. The Estimate corresponds to the **slope** of the predictor.
- **Standard Error**. It is the margin of error for each value. The lower, the better, because it would mean that there is little difference between values if the model was run multiple times.
- **t-value**. It is the result of dividing the Estimate by the SE. That is, it calculates the number of Standard Errors to obtain the Estimate. For example, if the estimate for probability is 0.03, and the SE is 0.001, the t-value would be 30. If the SE is 0.02, for example, the t-value would be 1,5. The higher, the better.
- **Pr (> |t|)**. P-value, basically. Assuming that H0 is the opposite to the hypothesis (there is no relation between the predictors and entropy), if the p-value is < 0.05, it would mean that H0 would be rejected and, thus, the hypothesis would be confirmed. That is, the lower from 0.05, the better. There are some symbols that also indicate the significance of the p-value, from highest signification (lower p-value) to the lowest (higher p-value): "***", "**", "*", ".", and " ".

Also, the model returns some general statistics regarding the estimations:

- The R-squared value is a value from 0 to 1 that indicates the percentage variation of entropy explained by the predictors.
- The F-statistic indicates the relationship between the predictors and entropy. The furthest from 1, the better. The distance from 1 will be meaningful according to the size of the dataset:
  - Large dataset: it does not need to be very far from 1 to indicate a strong relationship.
  - Small dataset: the value should be much higher than 1 to indicate a strong relationship.

By using the coefficients obtained in Figure 19, it is possible to define the model to predict the entropy (and, therefore, the difficulty of the example) by using the proposed predictors:

$$log(\texttt{ent}) = -20.3 + 4.6 \cdot \texttt{c.embed} - 0.006 \cdot \texttt{c.prob} + 0.033 \cdot \texttt{c.prob} \cdot \texttt{c.embed} \quad (36)$$

Also, in the results from the same figure, it is possible to determine the interaction between each predictor and entropy:

- When the values of the predictors are set to zero, the model shows the mean value for the log-entropy. Since data has been centered to its mean, the intercept becomes interpretable: when the values for the predictors are average (that is, -3.7 for similarity, and -88.7 for probability), the mean value for log-entropy is -20.3.

- Regarding definition **similarity**, it can be seen that it has a remarkable relationship with log-entropy, with a slope of 4.6, which means an increase in entropy of approximately a 4.6% each time the similarity increases 1 point. Taking into account the observations in the previous sections, in which it was observed that the higher the similarity, the higher the entropy, these results **agree** with the hypothesis.

- Regarding context sentence **probability**, it seems that the relationship with log-entropy is low, with a slope of -0.006, which means that, when the probability increases by one point, the entropy **decreases**. According to the results obtained in the previous experimentations, in which probability seemed to affect the entropy positively (that is, the higher the probability, the higher the entropy), these results do **not agree** with the hypothesis.

In these observations, the model also shows that there exists an interaction between the factors. More specifically, it looks like the behaviour of the probability of the context sentence might be affected by the average similarity of the senses of the word. In this sense, since the slope is very low, further analysis of the interaction is required in order to extract conclusions.

The histogram in Figure 20 shows the distribution of the (centered) averaged similarity values for each word in the dataset. The results show that the range of difficulty for most of the data is between -0.5 and 0.5.

As it has been mentioned earlier in this section, it looks like similarity affects the probability of the context sentence. Thus, it seems that, depending on the similarity value, the behaviour of probability changes. In order to be able to analyse this interaction, the interactions between entropy and probability, by taking into account similarity, will be plotted.

In Figure 21, data has been assigned three different values of similarity: low similarity, 0 similarity (medium) and high similarity, corresponding to the acceptable extreme values observed in Figure 20 (-0.5, 0, and 0.5, respectively). By doing so, is is possible to distinguish the behaviour of probability according to the difficulty of the word[24].

Furthermore, Figure 21 shows that, for **easy** words (that is, when the similarity is low), it seems that the higher the probability, the lower the entropy. That is, for easy words, the less informative the context sentence, the easier the example. In this case, thus, the

[24]Low similarity = easy word
High similarity = difficult word

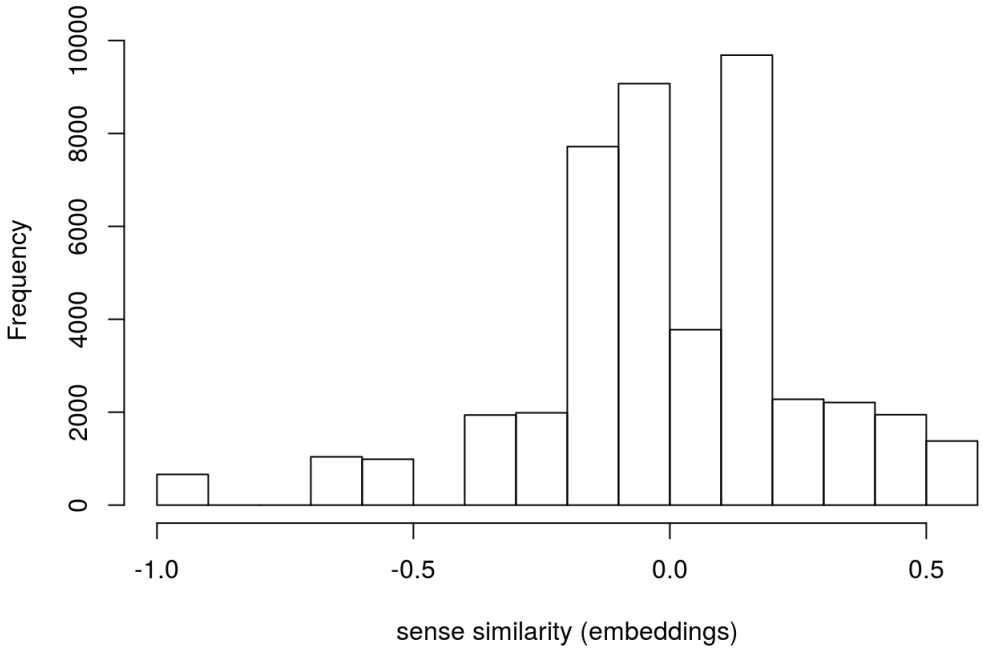


Figure 20: Histogram for distribution of similarity, by using centered data (`c.embeddings_sim`)

more *difficult* a context is[25], the easier to disambiguate, which, obviously, does not follow the hypothesis.

For **difficult** words (that is, when the similarity is high), the results show that the higher the probability, the higher the entropy. That is, the less informative the context sentence, the more difficult the example. Here, the hypothesis is indeed reflected.

As it can be seen, the observations show different behaviours depending on the difficulty of the word (i.e., its similarity value). It seems that when the word is **easy**, the information contained in the context sentence is not that relevant in the disambiguation task. In fact, it seems that the more informative −improbable− the sentence (or, even so, the more complex), the more difficult to disambiguate, leading to the intuition that **the information in the context is not very helpful** when disambiguating **easy** words.

This intuition may be explained by two facts:

- If the word is easy, it means that its senses may be easily distinguishable. Therefore, the context sentence does not need to be extremely informative, since the senses of

[25] According to the observations in the predictor, in which a difficult context sentence would be one with a high probability.

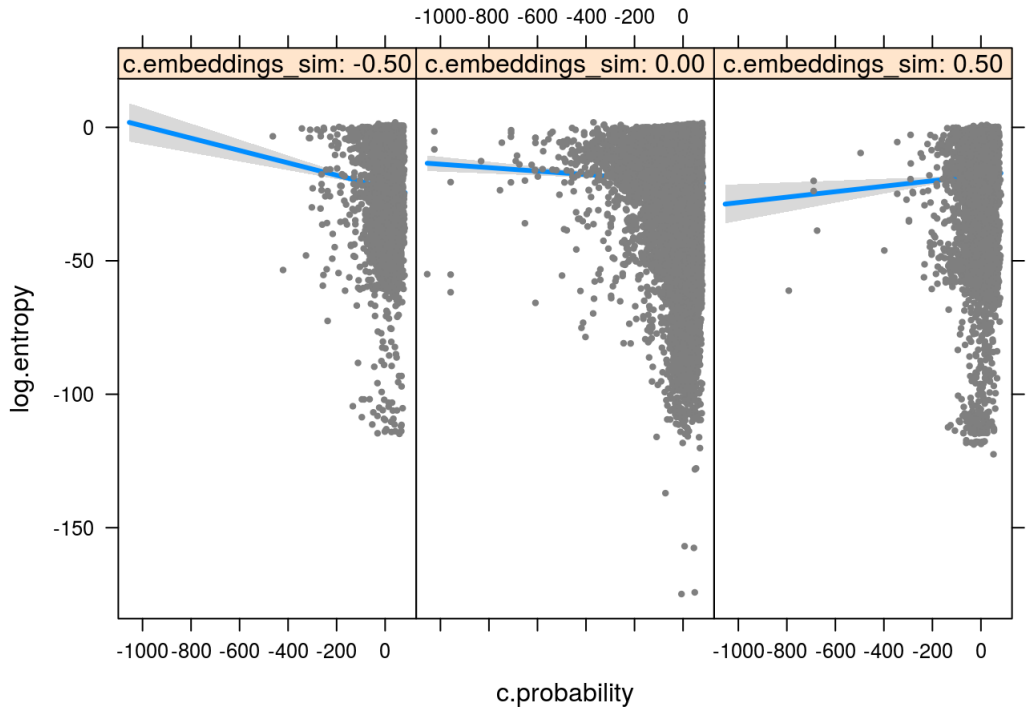


Figure 21: Plot for multifactorial analysis results, with easy words at the left (low similarity), medium words in the middle, and difficulty words at the right (high similarity).

the word may define very different situations and the probability of having a context sentence that may fit two or more senses is very low.

- The definition of probability related to the complexity of the context sentence to characterize difficulty is sometimes fuzzy. For example, common words can be enough to distinguish a sense from another. For example, in this example from Section 1.2.2:

  Word: *wait-v*

  Senses:

  - **1:** Stay in one place and anticipate or expect something.
  - **2:** Wait before acting.
  - **3:** Look forward to the probable occurrence of.
  - **4:** Serve as a waiter or waitress in a restaurant.

  Sentences:

- **1:** "Prudie, imagining herself **waiting** on *tables*, concurs that an appreciative gratuity is, indeed, preferable to repeated thank yous and now considers the problem solved". - **Sense 4**
- **2:** "Saatchi, by contrast, has kept a whole generation of artists from having to **wait** *tables*". - **Sense 4**

It can be seen that the words that help distinguish the sense (in italics) are very common words, as they are not very technical and widely used.

Therefore, for easy words, it is not necessary to have a very informative context and, thus, an improbable context sentence. In fact, if a context is too informative, or even technical, it may confuse the annotator, which may explain the results in Figure 21 and their divergence from the hypothesis.

When the word is **difficult**, the situation is different, since the senses for the word are not easy to distinguish between them and the annotators need to rely on another source of information that provides more hints about which sense to annotate. Thus, the context sentence should be more complex −and, thus, the probability should be lower− in order to provide more specific content, so as to be able to distinguish between senses that seem to be so fine-grained that are difficult to differentiate.

The finding of this interaction between the proposed predictors opens new avenues for future research lines.

5.3 Conclusions

In this section, the factors and measures that have been proposed in the previous sections to determine the difficulty of words and context sentences separately have been analysed together. The purpose to do so is to be able to design a model capable of predicting the difficulty of an example by using the predictors obtained and to analyse up to what extent these predictors behave in relation to entropy, which has been provided as the supervised measure to characterize the difficulty of an example.

The model proposed is based on **linear regression**, **log-scaling the explicative variable** (that is, entropy) and **centering the predictors** (that is, the similarity between the sense definitions of the word and the probability of the context sentences) in order to be able to better visualize and interpret the results.

The results show that there exists an interaction between the average similarity of senses and the probability of the context sentence. In order to better analyse this interaction, the results have been organized in a plot, distinguishing between low similarity, 0 similarity (medium similarity) and high similarity. That is, the results have been shown according to the difficulty of a word: easy words, medium words and difficult words, respectively.

This plot, Figure 21, has shown that, for **easy** words, the context sentence and its probability (and, thus, the information that it contains) are not very helpful in order to

disambiguate an example. Even so, it seems that a more probable context sentence (that is, a less informative/complex context sentence), makes the disambiguation task easier in these words, since it decreases the entropy value. This can be due to the fact that, in easy words, the senses are so distinguishable between them that is not necessary to have a highly explicative context sentence in order to be able to disambiguate it. This behaviour is **contrary** to the hypothesis.

However, for **difficult** words, the behaviour **does follow** the hypothesis, since the results show that an increase in the probability of the context sentences (and, hence, a lack of information) increases the entropy value of the example. This is probably due to the fact that, since the senses for difficult words are difficult to distinguish, the annotators have to rely on the context sentence in order to extract the maximum information so as to be able to choose an appropriate sense for the disambiguation task.

It seems that the probability of a context sentence is not very helpful in order to model the difficulty of examples, although it is true that when the similarity is high, the relation is more evident. However, and taking into account the results in the previous sections, in which the correlation between probability and entropy is not high, it is possible that the probability of the context sentence is not an appropriate measure to model the difficulty of context sentences, although it is a value that has signals of interaction with entropy and average similarity of senses.

In conclusion, the results have provided some signals of the effect of the predictors in the entropy value, being more evident the effect of the similarity of the words' sense definitions, and it has been observed the importance of the information in the context sentence in relation to the difficulty of the word, although the measure chosen to determine the difficulty of a context sentence, the probability, may not be very informative by itself, showing the need of searching for other measures that may be more descriptive.

6 Conclusions and final remarks

In this project, measures that determine difficulty in words and sentences in disambiguation tasks using WordNet have been presented, both for supervised and unsupervised methods.

The motivation comes from the fact that, nowadays, WSD systems do not achieve optimal results when using WordNet senses as word information, both for training and test data and, as López de Lacalle and Agirre (2015b)'s work proves, the fact of removing *difficult* sentences from training data improves the performance of automatic WSD systems. Also, it can open the possibility of developing alternative analyses for difficult examples.

The observations regarding disambiguation of examples by using WordNet senses have hinted that the two main factors that are related to the difficulty of annotation are the **words** (more specifically, their **senses**) and the **context sentences** to annotate, since they are the two main sources of information involved in a disambiguation task.

Several experiments have been performed in order to prove up to what extent the proposed unsupervised measures model difficulty for both factors, and the interactions of these measures with the overall difficulty of an example.

6.1 Contributions

The main contributions of this project are the following:

- Regarding supervised methods, it has been shown that it is possible to characterize difficulty of target words and sentences by using data from an annotation task. For words, **kappa agreement** has been proved to be a useful metric to characterize **word** difficulty, whereas **entropy** can characterize context sentence and example **difficulty**.
- It has also been shown that the kappa value is related to entropy, in the sense that the overall entropy for *easy* words is lower than for *difficult* words.
- **Similarity** between WordNet sense definitions has been proven to be an influential unsupervised factor when determining the difficulty of words, although it cannot be considered as a unique factor to establish difficulty, as the values for Pearson and Spearman have shown[26]. It has also been observed that embeddings-based similarity methods are more precise than overlap methods for this task.
- It has been proven that **probability** of occurrence is a factor to take into account in order to determine the difficulty of **context sentences**. Probability has been calculated by using a **language model**. The results, despite being noisy, have shown a **moderate correlation between probability and entropy**, although, like in

[26] Values for the embeddings-based similarity method.

the case of words, it is **not a strong correlation**, meaning that, again, probability **helps** determining difficulty of context sentences, but it is not the only factor to be taken into account.

- A model to determine the difficulty of examples has been designed and proposed. In this model, based on **linear regression** with multiple variables, the explicative variable (entropy) has been log-scaled and the predictors (the proposed unsupervised measures for both words and context sentences −similarity and probability, respectively), centered. The results have shown that the probability of the context sentence **is related** to the difficulty of disambiguation, with different outcomes depending on the difficulty of the word to disambiguate, modelled by **similarity**.

6.2 Further work and final considerations

Given that the final results have shown that the proposed factors have some, **but not total**, relationship with the difficulty of words and context sentences, the door has been opened to perform further analysis in order to identify **other factors** that may affect the difficulty of words and context sentences in order to be able to successfully predict difficulty without annotation data, such as the **part of speech** of the target word (Yarowsky and Florian 2002) or the interaction of the proposed factors with the **number of senses of the target word** (López de Lacalle and Agirre 2015b).

Regarding factor interactions, the model has shown that there exist interactions between factors, which suggests that predicting example difficulty is a **complex task**.

As it has been pointed out in Section 4.2, averaging the similarities for each sense definition in order to obtain a single value for each word may **mask problematic cases**. Future estimations of similarity for words may include alternatives such as considering the **range of similarities** for all the sense definitions of a word, or the use of a metric that takes into account the **distribution of similarity** of the sense definitions (in order to prevent high similarity values to be masked by lower similarities).

Furthermore, taking into account the results obtained in Section 4.3, it is especially necessary to define a new measure to model the difficulty of context sentences, since the results obtained in the analyses performed have shown that, although there are signals that the probability of a context sentence may help model its difficulty, the results have not been clear enough.

It can also be interesting to replicate the experiments in López de Lacalle and Agirre (2015b) to see if the results in WSD systems improve by removing the problematic examples obtained with the proposed model.

What becomes clear is that, since important considerations have been made in this project,

there is a lot of work pending in order to be able to predict the difficulty of examples in order to **improve WSD systems**.

References

Amazon Mechanical Turk. Wikipedia. URL: https://en.wikipedia.org/wiki/Amazon_Mechanical_Turk.

Bag-of-words model. Wikipedia. URL: https://en.wikipedia.org/wiki/Bag-of-words_model.

Brownlee, J. (2017). *A Gentle Introduction to the Bag-of-Words Model [blog entry]*. Machine Learning Mastery. URL: https://machinelearningmastery.com/gentle-introduction-bag-words-model/.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20 (1). DOI: 10.1177/001316446002000104.

Dawid, A. P. and A. M. Skene (1979). "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1). DOI: 10.2307/2346806.

Fellbaum, C., ed. (1998). *WordNet: an electronic lexical database*. MIT Press.

— (2005). "WordNet and Wordnets". In: *Encyclopedia of Language and Linguistics*. Ed. by K. Brown. Oxford: Elsevier, pp. 665–670. URL: http://wordnet.princeton.edu/.

Glen, S. (2013). *Covariance in Statistics*. StatisticsHowTo. URL: http://www.statisticshowto.com/covariance/.

— (2017). *Monotonic Relationship*. StatisticsHowTo. URL: http://www.statisticshowto.com/monotonic-relationship/.

Hovy, E. et al. (2006). "OntoNotes: The 90% Solution". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06. New York, New York: Association for Computational Linguistics, pp. 57–60. URL: http://dl.acm.org/citation.cfm?id=1614049.1614064.

Ide, N. et al. (2010). "The Manually Annotated Sub-corpus: A Community Resource for and by the People". In: *Proceedings of the ACL 2010 Conference Short Papers*. ACLShort '10. Uppsala, Sweden: Association for Computational Linguistics, pp. 68–73. URL: http://dl.acm.org/citation.cfm?id=1858842.1858855.

Koirala, C. and R. Y. Jee (2015). "Experimental analyses of the factors affecting the gradience in sentence difficulty judgments". In: *Proceedings of the 2015 EUROCALL Conference*, pp. 324–329.

López de Lacalle, O. and E. Agirre (2015a). "A Methodology for Word Sense Disambiguation at 90% based on large-scale CrowdSourcing". In: *SEM@NAACL-HLT*, pp. 61–70.

— (2015b). "Crowdsourced Word Sense Annotations and Difficult Words and Examples". In: *IWCS*. The Association for Computer Linguistics, pp. 94–100.

Lu, Xiaofei (2010). "Automatic analysis of syntactic complexity in second language writing". In: *International Journal of Corpus Linguistics* 15.4, pp. 474–496. DOI: http://dx.doi.org/10.1075/ijcl.15.4.02lu. URL: http://www.jbe-platform.com/content/journals/10.1075/ijcl.15.4.02lu.

Manning, C. D, P. Raghavan, and H. Schütze. *Tokenization*. URL: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html.

Martínez Alonso, H. et al. (2015). "Predicting word sense annotation agreement". In: *LSDSem@EMNLP*. Association for Computational Linguistics, pp. 89–94.

Miller, G. A., R. Beckwith, et al. (1990). "Introduction to WordNet: An On-line Lexical Database". In: *International Journal of Lexicography* 3 (4). DOI: 10.1093/ijl/3.4.235.

Miller, G. A. and C. Leacock (2000). "Lexical Representations for Sentence Processing". In: *Polysemy: Theoretical and Computational Approaches*. Ed. by Y. Ravin and C. Leacock. Oxford University Press, pp. 152–160.

Passonneau, R. J., C. F. Baker, et al. (2012). "The MASC Word Sense Corpus". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA).

Passonneau, R. J. and B. Carpenter (2012). *MASC word sense sentence corpus, crowdsourced subset*.

— (2014). "The Benefits of a Model of Annotation". In: *Transactions of the Association for Computational Linguistics* 2, pp. 311–326. URL: https://transacl.org/ojs/index.php/tacl/article/view/389.

Pearson Product-Moment Correlation. Laerd Statistics. URL: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

Pennington, J., R. Socher, and C. D. Manning (2014). "Glove: Global Vectors for Word Representation". In: *EMNLP*. Vol. 14, pp. 1532–1543. URL: https://nlp.stanford.edu/pubs/glove.pdf.

Perrone, C.S. (2013). *Machine Learning. Cosine Similarity for Vector Space Models (Part III)*. URL: http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/.

Spearman's Rank-Order Correlation. Laerd Statistics. URL: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php.

Standard Deviation Formulas. MathsIsFun. URL: https://www.mathsisfun.com/data/standard-deviation-formulas.html.

University, Princeton. *What is WordNet?* URL: https://wordnet.princeton.edu/.

Yarowsky, D. and R. Florian (2002). "Evaluating sense disambiguation across diverse parameter spaces". In: *Natural Language Engineering* 8 (4). DOI: 10.1017/S135132490200298X.

Zhong, Z. and H. T. Ng (2010). "It makes sense: A wide-coverage Word Sense Disambiguation system for free text". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 78–83.

Appendix I: Senses for words in the dataset

WordPos	SenseId	Definition	Examples
add-v	1	make an addition (to); join or combine or unite with others; increase the quality, quantity, size or scope of	We added two students to that dorm room. She added a personal note to her letter. Add insult to injury. Add some extra plates to the dinner table.
add-v	2	state or say further	'It doesn't matter,' he supplied.
add-v	3	bestow a quality on	Her presence lends a certain cachet to the company. The music added a lot to the play. She brings a special atmosphere to our meetings. This adds a light note to the program.
add-v	4	make an addition by combining numbers	Add 27 and 49, please!.
add-v	5	determine the sum of	Add all the people in this town to those of the neighboring town.
add-v	6	constitute an addition	This paper will add to her reputation.

Table 9: Senses and examples for ***add-v***

WordPos	SenseId	Definition	Examples
appear-v	1	give a certain impression or have a certain outward aspect	She seems to be sleeping. This appears to be a very difficult problem. This project looks fishy. They appeared like people who had not eaten or slept for a long time.
appear-v	2	come into sight or view	He suddenly appeared at the wedding. A new star appeared on the horizon.
appear-v	3	be issued or published	Did your latest book appear yet?. The new Woody Allen film hasn't come out yet.
appear-v	4	seem to be true, probable, or apparent	It seems that he is very gifted. It appears that the weather in California is very bad.
appear-v	5	come into being or existence, or appear on the scene	Then the computer came along and changed our lives. Homo sapiens appeared millions of years ago.
appear-v	6	appear as a character on stage or appear in a play, etc.	Gielgud appears briefly in this movie. She appeared in 'Hamlet' on the London stage.

appear-v	7	present oneself formally, as before a (judicial) authority	He had to appear in court last month. She appeared on several charges of theft.

Table 10: Senses and examples for ***appear-v***

WordPos	SenseId	Definition	Examples
ask-v	1	make a request or demand for something to somebody	She asked him for a loan.
ask-v	2	direct or put; seek an answer to	ask a question.
ask-v	3	consider obligatory; request and expect	We require our secretary to be on time. Aren't we asking too much of these children?. I expect my students to arrive in time for their lessons.
ask-v	4	address a question to and expect an answer from	Ask your teacher about trigonometry. The children asked me about their dead grandmother. I inquired about their special today. He had to ask directions several times.
ask-v	5	require as useful, just, or proper	It takes nerve to do what she did. success usually requires hard work. This job asks a lot of patience and skill. This position demands a lot of personal sacrifice. This dinner calls for a spectacular dessert. This intervention does not postulate a patient's consent.
ask-v	6	make a date	Has he asked you out yet? He asekd me to a dance.
ask-v	7	require or ask for as a price or condition	He is asking $200 for the table. The kidnappers are asking a million dollars in return for the release of their hostage.

Table 11: Senses and examples for ***ask-v***

WordPos	SenseId	Definition	Examples
board-n	1	a committee having supervisory powers	the board has seven members.

board-n	2	a stout length of sawn timber; made in a wide variety of sizes and used for many purposes	
board-n	3	a flat piece of material designed for a special purpose	he nailed boards across the windows.
board-n	4	food or meals in general	she sets a fine table. room and board.
board-n	5	a vertical surface on which information can be displayed to public view	
board-n	6	a table at which meals are served	he helped her clear the dining table. a feast was spread upon the board.
board-n	7	electrical device consisting of a flat insulated surface that contains switches and dials and meters for controlling other electrical devices	he checked the instrument panel. suddenly the board lit up like a Christmas tree.
board-n	8	a printed circuit that can be inserted into expansion slots in a computer to increase the computer's capabilities	
board-n	9	a flat portable surface (usually rectangular) designed for board games	he got out the board and set up the pieces.

Table 12: Senses and examples for ***board-n***

WordPos	**SenseId**	**Definition**	**Examples**
book-n	1	a written work or composition that has been published (printed on pages bound together)	I am reading a good book on economics.
book-n	2	physical objects consisting of a number of pages bound together	he used a large book as a doorstop.
book-n	3	a compilation of the known facts regarding something or someone	Al Smith used to say, 'Let's look at the record'. his name is in all the record books.

book-n	4	a written version of a play or other dramatic composition; used in preparing for a performance	
book-n	5	a record in which commercial accounts are recorded	they got a subpoena to examine our books.
book-n	6	a collection of playing cards satisfying the rules of a card game	
book-n	7	a collection of rules or prescribed standards on the basis of which decisions are made	they run things by the book around here.
book-n	8	the sacred writings of Islam revealed by God to the prophet Muhammad during his life at Mecca and Medina	
book-n	9	the sacred writings of the Christian religions	he went to carry the Word to the heathen.
book-n	10	a major division of a long written composition	the book of Isaiah.
book-n	11	a number of sheets (ticket or stamps etc.) bound together on one edge	he bought a book of stamps.

Table 13: Senses and examples for ***book-n***

WordPos	**SenseId**	**Definition**	**Examples**
color-n	1	a visual attribute of things that results from the light they emit or transmit or reflect	a white color is made up of many different wavelengths of light.
color-n	2	interest and variety and intensity	the Puritan Period was lacking in color. the characters were delineated with exceptional vividness.
color-n	3	the timbre of a musical sound	the recording fails to capture the true color of the original music.
color-n	4	a race with skin pigmentation different from the white race (especially Blacks)	

color-n	5	an outward or token appearance or form that is deliberately misleading	he hoped his claims would have a semblance of authenticity. he tried to give his falsehood the gloss of moral sanction. the situation soon took on a different color.
color-n	6	any material used for its color	she used a different color for the trim.
color-n	7	(physics) the characteristic of quarks that determines their role in the strong interaction	each flavor of quarks comes in three colors.
color-n	8	the appearance of objects (or light sources) described in terms of a person's perception of their hue and lightness (or brightness) and saturation	

Table 14: Senses and examples for ***color-n***

WordPos	**SenseId**	**Definition**	**Examples**
common-j	1	belonging to or participated in by a community as a whole; public	for the common good. common lands are set aside for use by all members of a community.
common-j	2	having no special distinction or quality; widely known or commonly encountered; average or ordinary or usual	the common man. a common sailor. the common cold. a common nuisance. followed common procedure. it is common knowledge that she lives alone. the common housefly. a common brand of soap.
common-j	3	common to or shared by two or more parties	a common friend. the mutual interests of management and labor.
common-j	4	commonly encountered	a common (or familiar) complaint. the usual greeting.
common-j	5	being or characteristic of or appropriate to everyday language	common parlance. a vernacular term. vernacular speakers. the vulgar tongue of the masses. the technical and vulgar names for an animal species.

common-j	6	of or associated with the great masses of people	the common people in those days suffered greatly. behavior that branded him as common. his square plebeian nose. a vulgar and objectionable person. the unwashed masses.
common-j	7	of low or inferior quality or value	of what coarse metal ye are molded- Shakespeare. produced...the common cloths used by the poorer population.
common-j	8	lacking refinement or cultivation or taste	he had coarse manners but a first-rate mind. behavior that branded him as common. an untutored and uncouth human being. an uncouth soldier–a real tough guy. appealing to the vulgar taste for violence. the vulgar display of the newly rich.
common-j	9	to be expected; standard	common decency.

Table 15: Senses and examples for ***common-j***

WordPos	SenseId	Definition	Examples
control-n	1	power to direct or determine	under control.
control-n	2	a relation of constraint of one entity (thing or person or group) by another	measures for the control of disease. they instituted controls over drinking on campus.
control-n	3	(physiology) regulation or maintenance of a function or action or reflex etc	the timing and control of his movements were unimpaired. he had lost control of his sphincters.
control-n	4	a standard against which other conditions can be compared in a scientific experiment	the control condition was inappropriate for the conclusions he wished to draw.
control-n	5	the activity of managing or exerting control over something	the control of the mob by the police was admirable.
control-n	6	the state that exists when one person or group has power over another	her apparent dominance of her husband was really her attempt to make him pay attention to her.
control-n	7	discipline in personal and social activities	he was a model of polite restraint. she never lost control of herself.

control-n	8	great skillfulness and knowledge of some subject or activity	a good command of French.
control-n	9	a mechanism that controls the operation of a machine	the speed controller on his turntable was not working properly. I turned the controls over to her.
control-n	10	a spiritual agency that is assumed to assist the medium during a seance	
control-n	11	the economic policy of controlling or limiting or curbing prices or wages etc.	they wanted to repeal all the legislation that imposed economic controls.

Table 16: Senses and examples for ***control-n***

WordPos	SenseId	Definition	Examples
date-n	1	the specified day of the month	what is the date today?.
date-n	2	a participant in a date	his date never stopped talking.
date-n	3	a meeting arranged in advance	she asked how to avoid kissing at the end of a date.
date-n	4	a particular but unspecified point in time	they hoped to get together at an early date.
date-n	5	the present	they are up to date. we haven't heard from them to date.
date-n	6	the particular day, month, or year (usually according to the Gregorian calendar) that an event occurred	he tried to memorizes all the dates for his history class.
date-n	7	a particular day specified as the time something happens	the date of the election is set by law.
date-n	8	sweet edible fruit of the date palm with a single long woody seed	

Table 17: Senses and examples for ***date-n***

WordPos	SenseId	Definition	Examples

fair-j	1	free from favoritism or self-interest or bias or deception; conforming with established standards or rules	a fair referee. fair deal. on a fair footing. a fair fight. by fair means or foul.
fair-j	2	not excessive or extreme	a fairish income. reasonable prices.
fair-j	3	very pleasing to the eye	my bonny lass. there's a bonny bay beyond. a comely face. young fair maidens.
fair-j	4	(of a baseball) hit between the foul lines	he hit a fair ball over the third base bag.
fair-j	5	lacking exceptional quality or ability	a novel of average merit. only a fair performance of the sonata. in fair health. the caliber of the students has gone from mediocre to above average. the performance was middling at best.
fair-j	6	attractively feminine	the fair sex.
fair-j	7	(of a manuscript) having few alterations or corrections	fair copy. a clean manuscript.
fair-j	8	gained or earned without cheating or stealing	an honest wage. an fair penny.
fair-j	9	free of clouds or rain	today will be fair and warm.
fair-j	10	(used of hair or skin) pale or light-colored	a fair complexion.

Table 18: Senses and examples for ***fair-j***

WordPos	**SenseId**	**Definition**	**Examples**
family-n	1	a social unit living together	he moved his family to Virginia. It was a good Christian household. I waited until the whole house was asleep. the teacher asked how many people made up his home. the family refused to accept his will.
family-n	2	primary social group; parents and children	he wanted to have a good job before starting a family.
family-n	3	a collection of things sharing a common attribute	there are two classes of detergents.
family-n	4	people descended from a common ancestor	his family has lived in Massachusetts since the Mayflower.

family-n	5	a person having kinship with another or others	he's kin. he's family.
family-n	6	(biology) a taxonomic group containing one or more genera	sharks belong to the fish family.
family-n	7	a loose affiliation of gangsters in charge of organized criminal activities	
family-n	8	an association of people who share common beliefs or activities	the message was addressed not just to employees but to every member of the company family. the church welcomed new members into its fellowship.

Table 19: Senses and examples for ***family-n***

WordPos	**SenseId**	**Definition**	**Examples**
find-v	1	come upon, as if by accident; meet with	We find this idea in Plato. I happened upon the most wonderful bakery not very far from here. She chanced upon an interesting book in the bookstore the other day.
find-v	2	discover or determine the existence, presence, or fact of	She detected high levels of lead in her drinking water. We found traces of lead in the paint.
find-v	3	come upon after searching; find the location of something that was missed or lost	Did you find your glasses?. I cannot find my gloves!.
find-v	4	establish after a calculation, investigation, experiment, survey, or study	find the product of two numbers. The physicist who found the elusive particle won the Nobel Prize.
find-v	5	come to believe on the basis of emotion, intuitions, or indefinite grounds	I feel that he doesn't like me. I find him to be obnoxious. I found the movie rather entertaining.
find-v	6	perceive or be contemporaneous with	We found Republicans winning the offices. You'll see a lot of cheating in this school. The 1960's saw the rebellion of the younger generation against established traditions. I want to see results.

find-v	7	get something or somebody for a specific purpose	I found this gadget that will serve as a bottle opener. I got hold of these tools to fix our plumbing. The chairman got hold of a secretary on Friday night to type the urgent letter.
find-v	8	make a discovery, make a new finding	Roentgen discovered X-rays. Physicists believe they found a new elementary particle.
find-v	9	make a discovery	She found that he had lied to her. The story is false, so far as I can discover.
find-v	10	obtain through effort or management	She found the time and energy to take care of her aging parents. We found the money to send our sons to college.
find-v	11	decide on and make a declaration about	find someone guilty.
find-v	12	receive a specified treatment (abstract)	These aspects of civilization do not find expression or receive an interpretation. His movie received a good review. I got nothing but trouble for my good intentions.
find-v	13	perceive oneself to be in a certain condition or place	I found myself in a difficult situation. When he woke up, he found himself in a hospital room.
find-v	14	get or find back; recover the use of	She regained control of herself. She found her voice and replied quickly.
find-v	15	succeed in reaching; arrive at	The arrow found its mark.
find-v	16	accept and make use of one's personality, abilities, and situation	My son went to Berkeley to find himself.

Table 20: Senses and examples for ***find-v***

WordPos	SenseId	Definition	Examples
fold-v	1	bend or lay so that one part covers the other	fold up the newspaper. turn up your collar.
fold-v	2	incorporate a food ingredient into a mixture by repeatedly turning it over without stirring or beating	Fold the egg whites into the batter.

fold-v	3	cease to operate or cause to cease operating	The owners decided to move and to close the factory. My business closes every night at 8 P.M.. close up the shop.
fold-v	4	confine in a fold, like sheep	
fold-v	5	become folded or folded up	The bed folds in a jiffy.

Table 21: Senses and examples for ***fold-v***

WordPos	SenseId	Definition	Examples
full-j	1	containing as much or as many as is possible or normal	a full glass. a sky full of stars. a full life. the auditorium was full to overflowing.
full-j	2	constituting the full quantity or extent; complete	an entire town devastated by an earthquake. gave full attention. a total failure.
full-j	3	complete in extent or degree and in every particular	a full game. a total eclipse. a total disaster.
full-j	4	filled to satisfaction with food or drink	a full stomach.
full-j	5	(of sound) having marked deepness and body	full tones. a full voice.
full-j	6	having the normally expected amount	gives full measure. gives good measure. a good mile from here.
full-j	7	being at a peak or culminating point	broad daylight. full summer.
full-j	8	having ample fabric	the current taste for wide trousers. a full skirt.

Table 22: Senses and examples for ***full-j***

WordPos	SenseId	Definition	Examples
help-v	1	give help or assistance; be of service	Everyone helped out during the earthquake. Can you help me carry this table?. She never helps around the house.
help-v	2	improve the condition of	These pills will help the patient.
help-v	3	be of use	This will help to prevent accidents.
help-v	4	abstain from doing; always used with a negative	I can't help myself–I have to smoke. She could not help watching the sad spectacle.

help-v	5	help to some food; help with food or drink	I served him three times, and after that he helped himself.
help-v	6	contribute to the furtherance of	This money will help the development of literacy in developing countries.
help-v	7	take or use	She helped herself to some of the office supplies.
help-v	8	improve; change for the better	New slipcovers will help the old living room furniture.

Table 23: Senses and examples for ***help-v***

WordPos	SenseId	Definition	Examples
high-j	1	greater than normal in degree or intensity or amount	a high temperature. a high price. the high point of his career. high risks. has high hopes. the river is high. he has a high opinion of himself.
high-j	2	(literal meaning) being at or having a relatively great or specific elevation or upward extension (sometimes used in combinations like 'knee-high')	a high mountain. high ceilings. high buildings. a high forehead. a high incline. a foot high.
high-j	3	standing above others in quality or position	people in high places. the high priest. eminent members of the community.
high-j	4	used of sounds and voices; high in pitch or frequency	
high-j	5	happy and excited and energetic	
high-j	6	(used of the smell of meat) smelling spoiled or tainted	
high-j	7	slightly and pleasantly intoxicated from alcohol or a drug (especially marijuana)	

Table 24: Senses and examples for ***high-j***

WordPos	SenseId	Definition	Examples
image-n	1	an iconic mental representation	her imagination forced images upon her too awful to contemplate.

image-n	2	(Jungian psychology) a personal facade that one presents to the world	a public image is as fragile as Humpty Dumpty.
image-n	3	a visual representation (of an object or scene or person or abstraction) produced on a surface	they showed us the pictures of their wedding. a movie is a series of images projected so rapidly that the eye integrates them.
image-n	4	a standard or typical example	he is the prototype of good breeding. he provided America with an image of the good father.
image-n	5	language used in a figurative or nonliteral sense	
image-n	6	someone who closely resembles a famous person (especially an actor)	he could be Gingrich's double. she's the very image of her mother.
image-n	7	(mathematics) the set of values of the dependent variable for which a function is defined	the image of f(x) = xˆ2 is the set of all non-negative real numbers if the domain of the function is the set of all real numbers.
image-n	8	the general impression that something (a person or organization or product) presents to the public	although her popular image was contrived it served to inspire music and pageantry. the company tried to project an altruistic image.
image-n	9	a representation of a person (especially in the form of sculpture)	the coin bears an effigy of Lincoln. the emperor's tomb had his image carved in stone.

Table 25: Senses and examples for ***image-n***

WordPos	**SenseId**	**Definition**	**Examples**
kill-v	1	cause to die; put to death, usually intentionally or knowingly	This man killed several people when he tried to rob a bank. The farmer killed a pig for the holidays.
kill-v	2	thwart the passage of	kill a motion. he shot down the student's proposal.
kill-v	3	end or extinguish by forceful means	Stamp out poverty!.
kill-v	4	be fatal	cigarettes kill. drunken driving kills.
kill-v	5	be the source of great pain for	These new shoes are killing me!.

kill-v	6	overwhelm with hilarity, pleasure, or admiration	The comedian was so funny, he was killing me!.
kill-v	7	hit with so much force as to make a return impossible, in racket games	She killed the ball.
kill-v	8	hit with great force	He killed the ball.
kill-v	9	deprive of life	AIDS has killed thousands in Africa.
kill-v	10	cause the death of, without intention	She was killed in the collision of three cars.
kill-v	11	drink down entirely	He downed three martinis before dinner. She killed a bottle of brandy that night. They popped a few beer after work.
kill-v	12	mark for deletion, rub off, or erase	kill these lines in the President's speech.
kill-v	13	tire out completely	The daily stress of her work is killing her.
kill-v	14	cause to cease operating	kill the engine.
kill-v	15	destroy a vitally essential quality of or in	Eating artichokes kills the taste of all other foods.

Table 26: Senses and examples for ***kill-v***

WordPos	SenseId	Definition	Examples
know-v	1	be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about	I know that the President lied to the people. I want to know who is winning the game!. I know it's time.
know-v	2	know how to do or perform something	She knows how to knit. Does your husband know how to cook?.
know-v	3	be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt	I know that I left the key on the table. Galileo knew that the earth moves around the sun.
know-v	4	be familiar or acquainted with a person or an object	She doesn't know this composer. Do you know my sister?. We know this movie. I know him under a different name. This flower is known as a Peruvian Lily.

know-v	5	have firsthand knowledge of states, situations, emotions, or sensations	I know the feeling!. have you ever known hunger?. I have lived a kind of hell when I was a drug addict. The holocaust survivors have lived a nightmare. I lived through two divorces.
know-v	6	accept (someone) to be what is claimed or accept his power and authority	The Crown Prince was acknowledged as the true heir to the throne. We do not recognize your gods.
know-v	7	have fixed in the mind	I know Latin. This student knows her irregular verbs. Do you know the poem well enough to recite it?.
know-v	8	have sexual intercourse with	This student sleeps with everyone in her dorm. Adam knew Eve. Were you ever intimate with this man?.
know-v	9	know the nature or character of	we all knew her as a big show-off.
know-v	10	be able to distinguish, recognize as being different	The child knows right from wrong.
know-v	11	perceive as familiar	I know this voice!.

Table 27: Senses and examples for ***know-v***

WordPos	SenseId	Definition	Examples
land-n	1	the land on which real estate is located	he built the house on land leased from the city.
land-n	2	material in the top layer of the surface of the earth in which plants can grow (especially with reference to its quality or use)	the land had never been plowed. good agricultural soil.
land-n	3	territory over which rule or control is exercised	his domain extended into Europe. he made it the law of the land.
land-n	4	the solid part of the earth's surface	the plane turned away from the sea and moved back over land. the earth shook for several minutes. he dropped the logs on the ground.
land-n	5	the territory occupied by a nation	he returned to the land of his birth. he visited several European countries.
land-n	6	a domain in which something is dominant	the untroubled kingdom of reason. a land of make-believe. the rise of the realm of cotton in the south.

land-n	7	extensive landed property (especially in the country) retained by the owner for his own use	the family owned a large estate on Long Island.
land-n	8	the people who live in a nation or country	a statement that sums up the nation's mood. the news was announced to the nation. the whole country worshipped him.
land-n	9	a politically organized body of people under a single government	the state has elected a new president. African nations. students who had come to the nation's capitol. the country's largest manufacturer. an industrialized land.
land-n	10		
land-n	11	agriculture considered as an occupation or way of life	farming is a strenuous life. there's no work on the land any more.

Table 28: Senses and examples for ***land-n***

WordPos	SenseId	Definition	Examples
late-j	1	being or occurring at an advanced period of time or after a usual or expected time	late evening. late 18th century. a late movie. took a late flight. had a late breakfast.
late-j	2	after the expected or usual time; delayed	a belated birthday card. I'm late for the plane. the train is late. tardy children are sent to the principal. always tardy in making dental appointments.
late-j	3	of the immediate past or just previous to the present time	a late development. their late quarrel. his recent trip to Africa. in recent months. a recent issue of the journal.
late-j	4	having died recently	her late husband.
late-j	5	of a later stage in the development of a language or literature; used especially of dead languages	Late Greek.
late-j	6	at or toward an end or late period or stage of development	the late phase of feudalism. a later symptom of the disease. later medical science could have saved the child.
late-j	7	(used especially of persons) of the immediate past	the former president. our late President is still very active. the previous occupant of the White House.

Table 29: Senses and examples for ***late-j***

WordPos	SenseId	Definition	Examples
level-n	1	a position on a scale of intensity or amount or quality	a moderate grade of intelligence. a high level of care is required. it is all a matter of degree.
level-n	2	a relative position or degree of value in a graded group	lumber of the highest grade.
level-n	3	a specific identifiable position in a continuum or series or especially in a process	a remarkable degree of frankness. at what stage are the social sciences?.
level-n	4	height above ground	the water reached ankle level. the pictures were at the same level.
level-n	5	indicator that establishes the horizontal when a bubble is centered in a tube of liquid	
level-n	6	a flat surface at right angles to a plumb line	park the car on the level.
level-n	7	an abstract place usually conceived as having depth	a good actor communicates on several levels. a simile has at least two layers of meaning. the mind functions on many strata simultaneously.
level-n	8	a structure consisting of a room or set of rooms at a single position along a vertical scale	what level is the office on?.

Table 30: Senses and examples for ***level-n***

WordPos	SenseId	Definition	Examples
life-n	1	a characteristic state or mode of living	social life. city life. real life.
life-n	2	the experience of being alive; the course of human events and activities	he could no longer cope with the complexities of life.
life-n	3	the course of existence of an individual; the actions and events that occur in living	he hoped for a new life in Australia. he wanted to live his own life without interference from others. get a life! he is trying to rebuild his life.

life-n	4	the condition of living or the state of being alive	while there's life there's hope. life depends on many chemical and physical processes.
life-n	5	the period during which something is functional (as between birth and death)	the battery had a short life. he lived a long and happy life.
life-n	6	the period between birth and the present time	I have known him all his life.
life-n	7	the period from the present until death	he appointed himself emperor for life.
life-n	8	a living person	his heroism saved a life.
life-n	9	animation and energy in action or expression	it was a heavy play and the actors tried in vain to give life to it.
life-n	10	living things collectively	the oceans are teeming with life.
life-n	11	the organic phenomenon that distinguishes living organisms from nonliving ones	there is no life on the moon.
life-n	12	an account of the series of events making up a person's life	
life-n	13	a motive for living	pottery was his life.
life-n	14	a prison term lasting as long as the prisoner lives	he got life for killing the guard.

Table 31: Senses and examples for ***life-n***

WordPos	SenseId	Definition	Examples
live-v	1	be an inhabitant of or reside in	People lived in Africa millions of years ago. The people inhabited the islands that are now deserted. this kind of fish dwells near the bottom of the ocean. deer are populating the woods.
live-v	2	lead a certain kind of life; live in a certain style	we had to live frugally after the war.
live-v	3	continue to live and avoid dying	We went without water and food for 3 days. These superstitions survive in the backwaters of America. The race car driver lived through several very serious accidents. how long can a person last without food and water? One crash victim died, the other lived.

live-v	4	support oneself	he could barely exist on such a low wage. Can you live on $2000 a month in New York City? . Many people in t a day.
live-v	5	have life, be alive	Our great leader is no more. My grandfather lived until the end of war.
live-v	6	have firsthand knowledge of states, situations, emotions, or sensations	I know the feeling!. have you ever known hunger?. I have lived a kind of hell when I was a drug addict. The holocaust survivors have lived a nightmare. I lived through two divorces.
live-v	7	pursue a positive and satisfying existence	You must accept yourself and others if you really want to live.

Table 32: Senses and examples for ***live-v***

WordPos	SenseId	Definition	Examples
long-j	1	primarily temporal sense; being or indicating a relatively great or greater than average duration or passage of time or a duration as specified	a long life. a long boring speech. a long time. a long friendship. a long game. long ago. an hour long.
long-j	2	primarily spatial sense; of relatively great or greater than average spatial extension or extension as specified	a long road. a long distance. contained many long words. ten miles long.
long-j	3	of relatively great height	a race of long gaunt men- Sherwood Anderson. looked out the long French windows.
long-j	4	good at remembering	a retentive mind. tenacious memory.
long-j	5	holding securities or commodities in expectation of a rise in prices	is long on coffee. a long position in gold.
long-j	6	(of speech sounds or syllables) of relatively long duration	the English vowel sounds in 'bate', 'beat', 'bite', 'boat', 'boot' are long.
long-j	7	involving substantial risk	long odds.

long-j	8	planning prudently for the future	large goals that required farsighted policies. took a long view of the geopolitical issues.
long-j	9	having or being more than normal or necessary	in long supply.

Table 33: Senses and examples for ***long-j***

WordPos	SenseId	Definition	Examples
lose-v	1	fail to keep or to maintain; cease to have, either physically or in an abstract sense	She lost her purse when she left it unattended on her seat.
lose-v	2	fail to win	We lost the battle but we won the war.
lose-v	3	suffer the loss of a person through death or removal	She lost her husband in the war. The couple that wanted to adopt the child lost her when the biological parents claimed her.
lose-v	4	place (something) where one cannot find it again	I misplaced my eyeglasses.
lose-v	5	miss from one's possessions; lose sight of	I've lost my glasses again!.
lose-v	6	allow to go out of sight	The detective lost the man he was shadowing after he had to stop at a red light.
lose-v	7	fail to make money in a business; make a loss or fail to profit	I lost thousands of dollars on that bad investment!. The company turned a loss after the first year.
lose-v	8	fail to get or obtain	I lost the opportunity to spend a year abroad.
lose-v	9	retreat	
lose-v	10	fail to perceive or to catch with the senses or the mind	I missed that remark. She missed his point. We lost part of what he said.
lose-v	11	be set at a disadvantage	This author really suffers in translation.

Table 34: Senses and examples for ***lose-v***

WordPos	SenseId	Definition	Examples
meet-v	1	come together	I'll probably see you at the meeting. How nice to see you again!.
meet-v	2	get together socially or for a specific purpose	

meet-v	3	be adjacent or come together	The lines converge at this point.
meet-v	4	fill or meet a want or need	
meet-v	5	satisfy a condition or restriction	Does this paper meet the requirements for the degree?.
meet-v	6	satisfy or fulfill	meet a need. this job doesn't match my dreams.
meet-v	7	collect in one place	We assembled in the church basement. Let's gather in the dining room.
meet-v	8	get to know; get acquainted with	I met this really handsome guy at a bar last night!. we met in Singapore.
meet-v	9	meet by design; be present at the arrival of	Can you meet me at the train station?.
meet-v	10	contend against an opponent in a sport, game, or battle	Princeton plays Yale this weekend. Charlie likes to play Mary.
meet-v	11	experience as a reaction	My proposal met with much opposition.
meet-v	12	undergo or suffer	meet a violent death. suffer a terrible fate.
meet-v	13	be in direct physical contact with; make contact	The two buildings touch. Their hands touched. The wire must not contact the metal cover. The surfaces contact at this point.

Table 35: Senses and examples for ***meet-v***

WordPos	SenseId	Definition	Examples
normal-j	1	conforming with or constituting a norm or standard or level or type or social norm; not abnormal	serve wine at normal room temperature. normal diplomatic relations. normal working hours. normal word order. normal curiosity. the normal course of events.
normal-j	2	in accordance with scientific laws	
normal-j	3	being approximately average or within certain limits in e.g. intelligence and development	a perfectly normal child. of normal intelligence. the most normal person I've ever met.
normal-j	4	forming a right angle	

Table 36: Senses and examples for ***normal-j***

WordPos	SenseId	Definition	Examples
number-n	1	the property possessed by a sum or total or indefinite quantity of units or individuals	the number of parameters is small. the figure was about a thousand.
number-n	2	a concept of quantity involving zero and units	every number has a unique position in the sequence.
number-n	3	a short performance that is part of a longer program	he did his act three times every evening. she had a catchy little routine. it was one of the best numbers he ever did.
number-n	4	the number is used in calling a particular telephone	he has an unlisted number.
number-n	5	a symbol used to represent a number	he learned to write the numerals before he went to school.
number-n	6	one of a series published periodically	she found an old issue of the magazine in her dentist's waiting room.
number-n	7	a select company of people	I hope to become one of their number before I die.
number-n	8	a numeral or string of numerals that is used for identification and may be attached to accounts, memberships, etc.	she refused to give them her Social Security number.
number-n	9	a clothing measurement	a number 13 shoe.
number-n	10	a numbered item in a series	take the number 2 to the main square, then change to the number 5.
number-n	11	the grammatical category for the forms of nouns and pronouns and verbs that are used depending on the number of entities involved (singular or dual or plural)	in English the subject and the verb must agree in number.
number-n	12	an item of clothing	she preferred the black nylon number. this sweater is an all-wool number.

Table 37: Senses and examples for ***number-n***

WordPos	SenseId	Definition	Examples

paper-n	1	a material made of cellulose pulp derived mainly from wood or rags or certain grasses	
paper-n	2	an essay (especially one written as an assignment)	he got an A on his composition.
paper-n	3	a daily or weekly publication on folded sheets; contains news and articles and advertisements	he read his newspaper at breakfast.
paper-n	4	a medium for written communication	the notion of an office running without paper is absurd.
paper-n	5	a scholarly article describing the results of observations or stating hypotheses	he has written many scientific papers.
paper-n	6	a business firm that publishes newspapers	Murdoch owns many newspapers.
paper-n	7	the physical object that is the product of a newspaper publisher	when it began to rain he covered his head with a newspaper.

Table 38: Senses and examples for ***paper-n***

WordPos	SenseId	Definition	Examples
particular-j	1	unique or specific to a person or thing or category	the particular demands of the job. has a particular preference for Chinese art. a peculiar bond of sympathy between them. an expression peculiar to Canadians. rights peculiar to the rich. the special features of a computer. my own special chair.
particular-j	2	separate and distinct from others of the same group or category	interested in one particular artist. a man who wishes to make a particular woman fall in love with him.
particular-j	3	surpassing what is common or usual or expected	he paid especial attention to her. exceptional kindness. a matter of particular and unusual importance. a special occasion. a special reason to confide in her. what's so special about the year 2000?.

particular-j	4	first and most important	his special interest is music. she gets special (or particular) satisfaction from her volunteer work.
particular-j	5	exacting especially about details	a finicky eater. fussy about clothes. very particular about how her food was prepared.
particular-j	6	providing specific details or circumstances	a particular description of the room.

Table 39: Senses and examples for ***particular-j***

WordPos	SenseId	Definition	Examples
poor-j	1	deserving or inciting pity	a hapless victim. miserable victims of war. the shabby room struck her as extraordinarily pathetic- Galsworthy. piteous appeals for help. pitiable homeless children. a pitiful fate. Oh, you poor thing. his poor distorted limbs. a wretched life.
poor-j	2	having little money or few possessions	deplored the gap between rich and poor countries. the proverbial poor artist living in a garret.
poor-j	3	characterized by or indicating poverty	the country had a poor economy. they lived in the poor section of town.
poor-j	4	lacking in quality or substances	a poor land. the area was poor in timber and coal. food poor in nutritive value. the food in the cafeteria was of poor quality.
poor-j	5	of insufficient quantity to meet a need	an inadequate income. a poor salary. money is short. on short rations. food is in short supply. short on experience. the jejune diets of the very poor.

Table 40: Senses and examples for ***poor-j***

WordPos	SenseId	Definition	Examples
read-v	1	interpret something that is written or printed	read the advertisement. Have you read Salman Rushdie?.
read-v	2	have or contain a certain wording or form	The passage reads as follows. What does the law say?.

read-v	3	look at, interpret, and say out loud something that is written or printed	The King will read the proclamation at noon.
read-v	4	obtain data from magnetic tapes	This dictionary can be read by the computer.
read-v	5	interpret the significance of, as of palms, tea leaves, intestines, the sky; also of human behavior	She read the sky and predicted rain. I can't read his strange behavior. The fortune teller read his fate in the crystal ball.
read-v	6	interpret something in a certain way; convey a particular meaning or impression	I read this address as a satire. How should I take this message?. You can't take credit for this!.
read-v	7	be a student of a certain subject	She is reading for the bar exam.
read-v	8	indicate a certain reading; of gauges and instruments	The thermometer showed thirteen degrees below zero. The gauge read 'empty'.
read-v	9	audition for a stage role by reading parts of a role	He is auditioning for 'Julius Caesar' at Stratford this year.
read-v	10	to hear and understand	I read you loud and clear!.
read-v	11	make sense of a language	She understands French. Can you read Greek?.

Table 41: Senses and examples for ***read-v***

WordPos	SenseId	Definition	Examples
say-v	1	express in words	He said that he wanted to marry her. tell me what is bothering you. state your opinion. state your name.
say-v	2	report or maintain	He alleged that he was the victim of a crime. He said it was too late to intervene in the war. The registrar says that I owe the school money.
say-v	3	express a supposition	Let us say that he did not tell the truth. Let's say you had a lot of money–what would you do?.
say-v	4	have or contain a certain wording or form	The passage reads as follows. What does the law say?.
say-v	5	give instructions to or direct somebody to do something with authority	I said to him to go home. She ordered him to do the shopping. The mother told the child to get dressed.

say-v	6	speak, pronounce, or utter in a certain way	She pronounces French words in a funny way. I cannot say 'zip wire'. Can the child sound out this complicated word?.
say-v	7	communicate or express nonverbally	What does this painting say?. Did his face say anything about how he felt?.
say-v	8	utter aloud	She said 'Hello' to everyone in the office.
say-v	9	state as one's opinion or judgement; declare	I say let's forget this whole business.
say-v	10	recite or repeat a fixed text	Say grace. She said her 'Hail Mary'.
say-v	11	indicate	The clock says noon.

Table 42: Senses and examples for ***say-v***

WordPos	SenseId	Definition	Examples
sense-n	1	a general conscious awareness	a sense of security. a sense of happiness. a sense of danger. a sense of self.
sense-n	2	the meaning of a word or expression; the way in which a word or expression or situation can be interpreted	the dictionary gave several senses for the word. in the best sense charity is really a duty. the signifier is linked to the signified.
sense-n	3	the faculty through which the external world is apprehended	in the dark he had to depend on touch and on his senses of smell and hearing.
sense-n	4	sound practical judgment	Common sense is not so common. he hasn't got the sense God gave little green apples. fortunately she had the good sense to run away.
sense-n	5	a natural appreciation or ability	a keen musical sense. a good sense of timing.

Table 43: Senses and examples for ***sense-n***

WordPos	SenseId	Definition	Examples
serve-v	1	serve a purpose, role, or function	The tree stump serves as a table. The female students served as a control group. This table would serve very well. His freedom served him well. The table functions as a desk.

serve-v	2	do duty or hold offices; serve in a specific function	He served as head of the department for three years. She served in Congress for two terms.
serve-v	3	contribute or conduce to	The scandal served to increase his popularity.
serve-v	4	be used by; as of a utility	The sewage plant served the neighboring communities. The garage served to shelter his horses.
serve-v	5	help to some food; help with food or drink	I served him three times, and after that he helped himself.
serve-v	6	provide (usually but not necessarily food)	We serve meals for the homeless. She dished out the soup at 8 P.M.. The entertainers served up a lively show.
serve-v	7	devote (part of) one's life or efforts to, as of countries, institutions, or ideas	She served the art of music. He served the church. serve the country.
serve-v	8	promote, benefit, or be useful or beneficial to	Art serves commerce. Their interests are served. The lake serves recreation. The President's wisdom has served the country well.
serve-v	9	spend time in prison or in a labor camp	He did six years for embezzlement.
serve-v	10	work for or be a servant to	May I serve you?. She attends the old lady in the wheelchair. Can you wait on our table, please?. Is a salesperson assisting you?. The minister served the King for many years.
serve-v	11	deliver a warrant or summons to someone	He was processed by the sheriff.
serve-v	12	be sufficient; be adequate, either in quality or quantity	A few words would answer. This car suits my purpose well. Will $100 do?. A 'B' grade doesn't suffice to get me into medical school. Nothing else will serve.
serve-v	13	do military service	She served in Vietnam. My sons never served, because they are short-sighted.
serve-v	14	mate with	male animals serve the females for breeding purposes.
serve-v	15	put the ball into play	It was Agassi's turn to serve.

Table 44: Senses and examples for ***serve-v***

WordPos	SenseId	Definition	Examples
show-v	1	give an exhibition of to an interested audience	She shows her dogs frequently. We will demo the new software in Washington.
show-v	2	establish the validity of something, as by an example, explanation or experiment	The experiment demonstrated the instability of the compound. The mathematician showed the validity of the conjecture.
show-v	3	provide evidence for	The blood test showed that he was the father. Her behavior testified to her incompetence.
show-v	4	make visible or noticeable	She showed her talent for cooking. Show me your etchings, please.
show-v	5	show in, or as in, a picture	This scene depicts country life. the face of the child is rendered with much tenderness in this painting.
show-v	6	give expression to	She showed her disappointment.
show-v	7	indicate a place, direction, person, or thing; either spatially or figuratively	I showed the customer the glove section. He pointed to the empty parking space. he indicated his opponents.
show-v	8	be or become visible or noticeable	His good upbringing really shows. The dirty side will show.
show-v	9	indicate a certain reading; of gauges and instruments	The thermometer showed thirteen degrees below zero. The gauge read 'empty'.
show-v	10	give evidence of, as of records	The diary shows his distress that evening.
show-v	11	take (someone) to their seats, as in theaters or auditoriums	The usher showed us to our seats.
show-v	12	finish third or better in a horse or dog race	he bet $2 on number six to show.

Table 45: Senses and examples for ***show-v***

WordPos	SenseId	Definition	Examples
suggest-v	1	make a proposal, declare a plan for something	the senator proposed to abolish the sales tax.
suggest-v	2	drop a hint; intimate by a hint	
suggest-v	3	imply as a possibility	The evidence suggests a need for more clarification.

suggest-v	4	call to mind	this remark evoked sadness.

Table 46: Senses and examples for ***suggest-v***

WordPos	SenseId	Definition	Examples
tell-v	1	express in words	He said that he wanted to marry her. tell me what is bothering you. state your opinion. state your name.
tell-v	2	let something be known	Tell them that you will be late.
tell-v	3	narrate or give a detailed account of	Tell what happened. The father told a story to his child.
tell-v	4	give instructions to or direct somebody to do something with authority	I said to him to go home. She ordered him to do the shopping. The mother told the child to get dressed.
tell-v	5	discern or comprehend	He could tell that she was unhappy.
tell-v	6	inform positively and with certainty and confidence	I tell you that man is a crook!.
tell-v	7	give evidence	he was telling on all his former colleague.
tell-v	8	mark as different	We distinguish several kinds of maple.

Table 47: Senses and examples for ***tell-v***

WordPos	SenseId	Definition	Examples
time-n	1	an instance or single occasion for some event	this time he succeeded. he called four times. he could do ten at a clip.
time-n	2	a period of time considered as a resource under your control and sufficient to accomplish something	take time to smell the roses. I didn't have time to finish. it took more than half my time.
time-n	3	an indefinite period (usually marked by specific attributes or activities)	he waited a long time. the time of year for planting. he was a great actor in his time.
time-n	4	a suitable moment	it is time to go.
time-n	5	the continuum of experience in which events pass from the future through the present to the past	
time-n	6	a person's experience on a particular occasion	he had a time holding back the tears. they had a good time together.

time-n	7	a reading of a point in time as given by a clock	do you know what time it is?. the time is 10 o'clock.
time-n	8	the fourth coordinate that is required (along with three spatial dimensions) to specify a physical event	
time-n	9	rhythm as given by division into parts of equal duration	
time-n	10	the period of time a prisoner is imprisoned	he served a prison term of 15 months. his sentence was 5 to 10 years. he is doing time in the county jail.

Table 48: Senses and examples for ***time-n***

WordPos	SenseId	Definition	Examples
wait-v	1	stay in one place and anticipate or expect something	I had to wait on line for an hour to get the tickets.
wait-v	2	wait before acting	the scientists held off announcing their results until they repeated the experiment.
wait-v	3	look forward to the probable occurrence of	We were expecting a visit from our relatives. She is looking to a promotion. he is waiting to be drafted.
wait-v	4	serve as a waiter or waitress in a restaurant	I'm waiting on tables at Maxim's.

Table 49: Senses and examples for ***wait-v***

WordPos	SenseId	Definition	Examples
way-n	1	how something is done or how it happens	her dignified manner. his rapid manner of talking. their nomadic mode of existence. in the characteristic New York style. a lonely way of life. in an abrasive fashion.
way-n	2	thing or person that acts to produce a particular effect or achieve an end	a means of control. an example is the best agency of instruction. the true way to success.
way-n	3	a line leading to a place or point	he looked the other direction. didn't know the way home.
way-n	4	the condition of things generally	that's the way it is. I felt the same way.

way-n	5	a course of conduct	the path of virtue. we went our separate ways. our paths in life led us apart. genius usually follows a revolutionary path.
way-n	6	any artifact consisting of a road or path affording passage from one place to another	he said he was looking for the way out.
way-n	7	a journey or passage	they are on the way.
way-n	8	space for movement	room to pass. make way for. hardly enough elbow room to turn around.
way-n	9	the property of distance in general	it's a long way to Moscow. he went a long ways.
way-n	10	doing as one pleases or chooses	if I had my way.
way-n	11	a general category of things; used in the expression 'in the way of'	they didn't have much in the way of clothing.
way-n	12	a portion of something divided into shares	they split the loot three ways.

Table 50: Senses and examples for ***way-n***

WordPos	SenseId	Definition	Examples
win-v	1	be the winner in a contest or competition; be victorious	He won the Gold Medal in skating. Our home team won. Win the game.
win-v	2	win something through one's efforts	I acquired a passing knowledge of Chinese. Gain an understanding of international finance.
win-v	3	obtain advantages, such as points, etc.	The home team was gaining ground. After defeating the Knicks, the Blazers pulled ahead of the Lakers in the battle for the number-one playoff berth in the Western Conference.
win-v	4	attain success or reach a desired goal	The enterprise succeeded. We succeeded in getting tickets to the show. she struggled to overcome her handicap and won.

Table 51: Senses and examples for ***win-v***

WordPos	SenseId	Definition	Examples
window-n	1	a framework of wood or metal that contains a glass windowpane and is built into a wall or roof to admit light or air	
window-n	2	a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened	
window-n	3	a transparent panel (as of an envelope) inserted in an otherwise opaque material	
window-n	4	an opening that resembles a window in appearance or function	he could see them through a window in the trees.
window-n	5	the time period that is considered best for starting or finishing something	the expanded window will give us time to catch the thieves. they had a window of less than an hour when an attack would have succeeded.
window-n	6	a pane of glass in a window	the ball shattered the window.
window-n	7	an opening in a wall or screen that admits light and air and through which customers can be served	he stuck his head in the window.
window-n	8	(computer science) a rectangular part of a computer screen that contains a display different from the rest of the screen	

Table 52: Senses and examples for ***window-n***

WordPos	SenseId	Definition	Examples
work-n	1	activity directed toward making or doing something	she checked several points needing further work.

work-n	2	a product produced or accomplished through the effort or activity or agency of a person or thing	it is not regarded as one of his more memorable works. the symphony was hailed as an ingenious work. he was indebted to the pioneering work of John Dewey. the work of an active imagination. erosion is the work of wind or water over time.
work-n	3	the occupation for which you are paid	he is looking for employment. a lot of people are out of work.
work-n	4	applying the mind to learning and understanding a subject (especially by reading)	mastering a second language requires a lot of work. no schools offer graduate study in interior design.
work-n	5	(physics) a manifestation of energy; the transfer of energy from one physical system to another expressed as the product of a force and the distance through which it moves a body in the direction of that force	work equals force times distance.
work-n	6	a place where work is done	he arrived at work early today.
work-n	7	the total output of a writer or artist (or a substantial part of it)	he studied the entire Wagnerian oeuvre. Picasso's work can be divided into periods.

Table 53: Senses and examples for ***work-n***

Appendix II: Values for supervised and unsupervised measures

The following table shows, for each word in the dataset, the values for the **supervised** measure (**kappa**) and the values for the **unsupervised** measures (**similarity** using **overlap** methods and similarity using **embedding-based** methods) to determine difficulty in **words**:

word	overlap_avg_sim	embeddings_avg_sim	kappa
add-v	0.516	3.64	0.64
appear-v	0.901	3.88	0.56
ask-v	0.849	4.21	0.83
board-n	0.853	3.636	0.93
book-n	1.258	3.743	0.35
color-n	0.62	3.773	0.26
common-j	1.111	3.56	0.81
control-n	0.875	3.92	0.73
date-n	1.044	3.852	0.49
fair-j	0.817	2.814	0.92
family-n	0.74	3.721	0.68
find-v	0.571	3.851	0.73
fold-v	0.372	3.762	0.37
full-j	0.462	3.704	0.65
help-v	0.658	3.895	0.37
high-j	0.955	3.483	0.77
image-n	1.014	3.858	0.68
kill-v	0.31	3.329	0.66
know-v	0.683	3.971	0.13
land-n	0.85	3.923	0.77
late-j	1.241	4.03	0.92
level-n	1.214	3.901	0.24
life-n	0.933	3.908	0.81
live-v	0.357	3.567	0.87
long-j	0.621	3.503	0.76
lose-v	0.692	3.549	0.72
meet-v	0.574	3.551	0.78
normal-j	0.426	3.582	0.82
number-n	1.167	3.713	0.64
paper-n	0.599	3.574	0.76
particular-j	0.735	3.868	0.71
poor-j	0.608	3.177	0.75
read-v	0.729	3.77	0.9

say-v	0.602	3.116	0.47
sense-n	0.518	3.702	0.96
serve-v	0.637	3.563	0.72
show-v	0.683	3.649	0.69
suggest-v	0.833	3.677	0.55
tell-v	0.229	3.408	0.69
time-n	0.751	4.048	0.83
wait-v	0.167	3.578	0.56
way-n	0.715	3.864	0.56
win-v	0.182	3.678	0.39
window-n	1.445	4.176	0.63
work-n	0.879	4.246	0.57

Table 54: Values for the supervised and unsupervised measures to determine difficulty in words, for each word in the dataset