

# What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage

Annotation Guidebook by Christian M. Meyer and Iryna Gurevych



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT



UBIQUITOUS  
KNOWLEDGE  
PROCESSING

Version 1.0  
17.05.2010

---

With this annotation guidebook, you receive a spreadsheet that can be opened in Microsoft Excel or OpenOffice (please contact the authors in case of any problems). In the spreadsheet, you'll find a list of lemmas – one lemma per row. Each lemma is accompanied with two word senses that are represented by an id, a sense gloss (i.e., a definition text), and maybe some example sentences (written in quotes, where available). The word senses on the left hand side are taken from the English WordNet, the word senses on the right hand side from the English Wiktionary. For each lemma, all encoded word senses from WordNet are paired with all encoded word senses in Wiktionary in an ordered manner (i.e., all combinations are shown).

In between you'll find a column "Same Sense", which is to be annotated. Your task is to make a binary decision for each pair: Write a

- 0 if both word senses are different, or
- 1 if both word senses represent the same or a highly similar meaning

into the "Same Sense" column. The result will be a word sense alignment of WordNet and Wiktionary on a sample of lemmas.

*Examples:*

Lemma	WordNet	Same Sense	Wiktionary	Comment
<b>intersect</b>	meet at a point	1	to cross; to cut	<i>clearly the same</i>
<b>intersect</b>	meet at a point	0	(mathematics) two sets intersect if and only if they have at least one common element	<i>Wiktionary sense is domain specific</i>

Lemma	WordNet	Same Sense	Wiktionary	Comment
<b>people</b>	the body of citizens of a state or country; "the Spanish people"	1	("plural" peoples) Persons forming or belonging to a particular group, such as a nation, class, ethnic group, country, family, etc; folk; community.	<i>although not every item is mentioned in WordNet (such as class), one could consider the pair as highly similar</i>
<b>people</b>	(plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience"	1	(used as plural of person); a body of human beings considered generally or collectively; a group of two or more persons.	
<b>people</b>	(plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience"	0	One's colleagues or employees.	<i>The Wiktionary sense is a variant of the WordNet sense. Additionally, the previous sense fits nicely</i>
<b>people</b>	...	...	...	

Lemma	WordNet	Same Sense	Wiktionary	Comment
week	any period of seven consecutive days; "it rained for a week"	1	A period of seven days.	
week	any period of seven consecutive days; "it rained for a week"	0	The seven days beginning with Sunday or Monday.	
week	any period of seven consecutive days; "it rained for a week"	0	A subdivision of the month into longer periods of work days punctuated by shorter weekend periods of days for markets, rest, or religious observation such as a sabbath.	
week	hours or days of work in a calendar week; "they worked a 40-hour week"	0	A period of seven days.	
week	hours or days of work in a calendar week; "they worked a 40-hour week"	0	The seven days beginning with Sunday or Monday.	
week	hours or days of work in a calendar week; "they worked a 40-hour week"	0	A subdivision of the month into longer periods of work days punctuated by shorter weekend periods of days for markets, rest, or religious observation such as a sabbath.	<i>Somehow similar, but WordNet clearly focuses on the (western) calendar week as a reference, while Wiktionary regards other calendar systems</i>
week	a period of seven consecutive days starting on Sunday	0	A period of seven days.	<i>There is a better one...</i>
week	a period of seven consecutive days starting on Sunday	1	The seven days beginning with Sunday or Monday.	
week	a period of seven consecutive days starting on Sunday	0	A subdivision of the month into longer periods of work days punctuated by shorter weekend periods of days for markets, rest, or religious observation such as a sabbath.	

*Please Note:*

- Do not imitate an algorithm, rely on your intuition.
- For each lemma, multiple "1"s are possible.
- Sometimes no "1" is suitable for a lemma, i.e. the word senses are complementary
- If one of the word senses covers only a small part of the other word sense and there is a better correspondence available, this "smaller" word sense is considered a variant and should be annotated with "0".
- It is possible that one of the word senses seems to be "wrong", i.e. there is no such meaning for this lemma, the sense definition is erroneous, or not exhaustive enough to clearly separate the sense from others. This fact should not be taken into account, i.e. take the word sense for granted and focus on the decision whether there is a corresponding word sense in the other resource.

- 
- Sometimes the glosses are trimmed or normalized by JWKTl and thus do not clearly describe the word sense. It is allowed to consult Wiktionary for the original gloss. You can easily query the Wiktionary by typing <http://en.wiktionary.org/wiki/<LEMMA>> in your browser.
  - The same applies to WordNet, where the word sense definitions might not be enough to infer the meaning. You are allowed to consult the Web frontend of WordNet at <http://wordnetweb.princeton.edu/perl/webwn>
  - As the task gets monotonous at some point, you should consider making breaks during the annotation process. Depending on your personal speed, it should take about 2–3 hours
  - Please do not change the order of the rows/cols or their IDs. We'll need this information for further processing. Of course, you are allowed to change the size of the columns to exactly fit on your screen...
  - If you feel like explaining a certain decision, you can use the comment column. This is optional; **you do not have to leave a comment for every pair!**

After you finished your annotations, please fill out the last row with the annotation time needed in minutes. There is no correct solution for this task, so please, always judge the sense pairs based on your own understanding. We will measure the inter-annotator agreement of the submissions to find out, how often the participants agree or disagree.