The Annotation process

For the preparation of BECL, noun-sense pairs were annotated by four native speakers of Canadian English. The annotation phases as well as the process of annotating itself are described below.

1. **Phases**

Annotation in group

Noun senses are being annotated starting from the top (order of noun ID). If not mentioned otherwise, only the first sense of a noun is annotated. Majority vote wins (after discussion). In case of tie, annotated as *don't know*

No.	Starting date	Description	Annotators
1	7.5.13	TRAINING; Annotation in group with guidance.	LS, FW, MJ, MD
2	8.5.13	TRAINING; Annotation in group with guidance	LS, FW
3	9.5.13	TRAINING; Annotation in group with some guidance	LS, FW, MJ
4	10.5.13 – 13.05.13	Annotation in group almost without guidance. First time annotators are more or less on their own. Phase 4 and 7 is done in parallel to phases 5 and 6. So half time they discuss nouns in group and the other half they are on their own.	LS, FW, MJ, MD
7	14.05.13	TST no longer present. Annotators are on their own. LAST ID annotated in group 4945	LS, FW, MJ, MD

Annotation on their own (no discussion of problematic nouns!)

No.	Starting date	Noun ID to ID	Description	Annotator
5.1	10.5.13 – 27.5.13	38987 - 40968	500 nouns given to both annotators	LS, MD
6.1	10.5.13 - 27.5.13	40972 - 42863	500 nouns given to both annotators	FW, MJ

5.2	10.5.13 – 27.5.13	38987 - 40968	500 nouns given to both annotators	FW, MJ
6.2	10.5.13 – 27.5.13	40972 - 42863	500 nouns given to both annotators	LS, MD
8	28.5.13 – 18.06.13	26827 - 38983	3000 nouns given to both annotators	LS, MJ
9	28.5.13 – 18.06.13	14724 - 26825	3000 nouns given to both annotators	FW, MD
10	18.6.13 - 05.07.13	4948 - 14723	2451 nouns given to all annotators, BUT this time all sense descriptions are deleted. The annotators just see the noun. The idea is to find dual-life nouns. Data is not contained in merged Master-Annotationfile	LS, FW, MD, MJ
11	09.07 – 03.08		Nouns from phase 8 are used. Annotators are told to annotate the senses 2, 3, 4 in WordNet of these nouns (if available)	LS, MJ
12	09.07 -		Nouns from phase 9 are used. Annotators are told to annotate the senses 2, 3, 4 in WordNet of these nouns (if available)	FW, MD
13	13.08 -	4948 - 14723	Nouns from phase 10 are annotated again. This time with description. Annotators are told to annotate all senses of a noun. (It was planed that MD also annotates these nouns, but had	FW, ~~MD~~

			to pass on this task due to lack of time)	
abstract_14	02.08 -	26827 - 38983	First four senses of 3000 nouns from phase 8 + 11 are used LS and MJ got their previously annotated files. Annotators are told to decide whether a noun is abstract or concrete. See external documents for discussion and criteria.	LS, MJ

2. **Speed of annotation**

Times measured on 13.5.2013 within group annotation (Phase 4)

Start time	**ID noun**	**End time**	**ID noun**	**Time in minutes**	**# nouns**	**Nouns/h**
10:12	1514	11:18	1697	66	44	40
11:19	1698	12:04	1818	45	29	39
12:11	1832	12:38	1946	27	30	67

Mean = 45 nouns/hour

3. **Remarks on disagreements**

Some clarification concerning Test II:

If noun does not have plural form -> *not applicable*

If noun does have plural form, but is ungrammatical in Test II.1 -> *NO*

Only in cases where noun is not quantifiable! Usually *not applicable* or *yes*

If *plural only noun* is grammatical in Test II.1 than Test II.2 is set to NOT APPLICABLE, as no singular form can be established for second sentence.

Sources of disagreement during IAA

Possible combinations of values in tests 2.1 and 2.2 leading to disagreement

example	A	B	remarks	
Tracking **Extortion** **implementation**	Not applicable (no plural)	Has plural and is not a kind-reading (fully countable)	Decision whether an event of this is considered/can be counted or not?! (mostly nominalizations obviously) Could depend on the (bad) description of WorldNet or the bad definition of how to handle events…	
Emerald **Shit** **Reputation** **infatuation**	Not applicable (no plural)	Has plural and is not a kind-reading	A says is mass. B says it is dual life! (so A assumes a second entry in WordNet later on?	Should reflect itself also in test 3.1 (indef usage only if dual life)
Celluloid **Celery** **Heroin** **Bleach** **copper**	Not applicable (no plural) (typical mass-noun)	Has kind reading (plural possible) (mass-noun with kind-reading through plural usage)	Quite arbitrary decision? It is not a question of semantics, but of ?!?!?. morphology?/ Frequency of noun in daily usage?	It is obvious what is happening here, but how to prevent this?!
meantime **Manipulation** **Weakening** **Workmanship**	Not applicable (no plural)	Has plural, but is ungrammatical in context of 'more' ('No')	Sometimes, it is a simple case of wrongly chosen values for test 2.1 (should be 'not applicable' but 'no' chosen instead). But there is more to this in some cases	

Comments of annotators on these disagreements:

<u>Dual life/plural kind-reading cases:</u>

"After looking at the Google document and talking about why there may have been disagreement, we were not able to come up with any fail-safe solutions for prevention. For the dual life cases the problem often seems to be a lack of precise definitions in Word net, as you suggest. We all have a good grasp on the no/not applicable distinction in 2.1 and feel that any error here are pure mistakes and not the result of misunderstanding. The plural

kind-reading cases seem to be based on experience/world knowledge, something which obviously differs between us."

Events:

"Disambiguating between events and process/state readings of verb is also a source of difference between us. Like the Dual life/plural kind-reading cases word-net definitions are a part of the problem as is, to some extent, the experience/world knowledge issue. We agreed to make use of the additional result/process/act columns as well as the comments column to provide more detail about the reasons for our decisions."

From one MJ:

"In a group meeting (just Matt, Fiona and me as Lisa is away presenting at conference) we discussed some of the problems that are happening the inter-rater reliability. For test 2.1 is seems like the source of the problem might be differences between us in whether or not a 'kind' reading is available for certain mass nouns. Some of us appear to be far more willing to allow 'kind' readings than others. In fact there is often a 50/50 split between us when such nouns are discussed in the group sessions. Obviously our responses to test 2.2 would best indicate which of use allow 'kind' readings. We noticed that you have not yet calculated reliability for this test. Once you have had the opportunity to the response to 2.1 might be a lot less confusing."

"I don't have too many overall comments other than those I have mentioned before. As I mention previously I think that the "dual-life" issue might be made simpler were a distinction made between the classic dual life cases (like "cake") and the ones that appear as dual life due to an act/process ambiguity. The latter case is (most often) related to nominalization and therefore might be filtered out with some of the aberrant processes that may be occurring with those. This might make the overall dual life picture seem like less of a conundrum.

In terms of the contrast between the description-less files and the files with descriptions I noticed a few things. Generally, it is much faster to annotate without the descriptions. I think this is because when there are no description it is not necessary to spend time teasing apart the different senses listed in wordnet. When the descriptions are present this seemed to take a large portion of the time. However, when there are no descriptions sometimes it is difficult (if not impossible) to decide on the meaning of a word without the description. This I think could lead to a lower rate of inter rater reliability as we may not be annotating the sames sense of the word. Because of this there also may be no way to determine if differences between us are the result of different countability judgements or merely due to different intuitions about which senses of the word is the most general and therefore the best sense to annotate."

"Regarding the question at hand, I think that in certain contexts 'secretion' is fine as a substance-mass noun. Taking from the context in the sentence you provide, a medical context in which several glads are compared for the rates at which they secrete different substances would work. Like many verbal nationalizations it often sounds more natural to use the corresponding verb in comparative structures (eg. ...secretes more than...). This I

feel might make judgments around such nouns a little cloudy, due to their rarity, but it doesn't mean they are unacceptable in my judgement.“

4. Adjudication process

Additional opinion on cases of disagreement was given by one annotator. The adjudicator can see how the two previous annotators annotated.

In case of a nominalization, the annotator has to annotate the reading he/she assumes during annotation. In the best case, he/she tries to stick to the assumed reading of the original annotators.

Adjudicator is also told to watch out to annotate test II.1 correctly: NA if noun has no plural; NO if noun has plural but is ungrammatical in test context.

MJ started on 30.10.2013 to adjudicate 2,001 cases of disagreement in the total of 6,429 cases annotated by MD and FW.

LS started in December 2013 to adjudicate 1,820 cases of disagreement in the annotation by LS and MJ.