The Bochum English Countability Lexicon, Distribution 2.0

Tibor Kiss, Halima Husic, Francis Jeffry Pelletier

{tibor, husic}@linguistics.rub.de, francisp@ualberta.ca

15.10.2014

1. Introduction: What is the Bochum English Countability Lexicon?

Many previous approaches to analysing the count/mass distinction made the following two assumptions, either implicitly or explicitly:

- The count/mass distinction can be analysed in binary fashion.
- The count/mass distinction can be analysed at the level of the noun lemma.

In providing the Bochum English Countability Lexicon (BECL for short), we challenge both assumptions. Starting with the second assumption, the BECL provides analyses not at the level of the noun lemma, but at the level of the noun sense. The noun senses have been taken from WordNet (searches for senses can be carried out under http://wordnetweb.princeton.edu/perl/webwn). Currently, we provide a maximum number of four senses per noun. This is not because we assume that a noun must not have more than four senses, but because the noun-sense pairs have been manually annotated, which is a time consuming task that did not allow us to annotate all senses of all nouns included in the lexicon. We envisage however that future releases will contain complete annotations for a larger subset of noun-sense pairs. By annotating noun-sense pairs, we are able to distinguish between nouns with different senses that fall into the same countability class, and nouns with different senses that fall into different countability classes.

We have not yet said what we consider to be a countability class, i.e. what our pick with regard to the first assumption is. In distributing the BECL, we make use of a bottom up strategy to identify possibly countability classes. In fact, as will become clear immediately, we argue for a logical class space of 729 classes, from which, however, only 18 classes are filled in the current stage of the annotation.

The basic idea behind the bottom up approach is to let annotators answer questions about possible insertions of nouns with particular senses into predefined pattern. A more complete picture of the methodology is provided in Kiss, Pelletier, and Stadtfeld (2014), which forms part of this distribution, so we will introduce the idea here only briefly.

In total, we assume that there are three basic patterns into which a noun can be inserted to get information about its countability class membership. The three patterns are divided into pairs of two, where each pattern may be answered affirmatively (such as *grammatical),* negatively (such as *ungrammatical),* or finally by indicating that the pattern is not applicable for the noun.

As an illustration for the three patterns consider the first pattern:

(1) Does inserting *noun#x* (where #x is the sense *x* of the *noun)* into **NP_1 VERB more *noun*$_{sg}$ than NP_2** lead to grammaticality, ungrammaticality, or is it not possible.

The noun *car* with its first sense (car#1, i.e. *a motor vehicle with four wheels; usually propelled by an internal combustion engine)* leads to ungrammaticality, i.e. a negative answer, inserting nouns like *fruitcake#2* and *lingerie#1* leads to grammaticality. Pluralia tantum (i.e. plural-only nouns, i.e., nouns without a singular form) such as *goggles#1* cannot be inserted into the first test pattern, leading to *not applicable.*

If the insertion of the noun resulted in a grammatical sentence in the first step of this test pattern, the annotator has to decide whether the comparison in the constructed sentence is based on the number of entities (e.g., pieces in the case of *lingerie#1),* or on a different kind of measurement (e.g., mass/volume in the case of *fruitcake#2).* This second step employs the experimental results provided in (Barner and Snedeker, 2005 and Bale and Barner, 2009). So here we take *number of entities* to be the affirmative answer, a different mode of measurement to be the negative answer, and in case the first pattern leads to ungrammaticality or is simply not applicable, the answer to the second question will also be *not applicable.*

If the answers are taken as features (affirmative, negative, not applicable), we allow a feature space of 3^6 = 729 possibly classes, but since certain answers are interdependent (as answering *not applicable* to the first question of a pair will only allow answering *not applicable* to the second question of the pair), the space is actually reduced.

It is crucial to understand that annotators do not answer questions about countability directly. Instead they have answered questions about possible insertions of nouns with certain senses into fixed patterns. For further questions concerning the test patterns, we would like to direct the reader to Kiss, Pelletier, and Stadtfeld (2014), which is included in this distribution.

So, for a user, the BECL provides a feature space, which is a subset from 729 possible answers to our patterns. How does the user arrive at a countability class for a noun, then?

2. Defining classes in BECL

In total, the raw version of the BECL contains annotations for about 15,000 noun-sense pairs. The nouns have been annotated in different stages by maximally four annotators, first in a group stage, where all annotators were trained by Tobias Stadtfeld, and reached a conclusion about an annotation as a group, and then in eleven further annotation stages, where each noun-sense pair has been annotated by at least two annotators.

We have employed the following strategy for binning the annotations into classes, which led to a total of 18 classes, comprising 10,667 annotated noun-sense pairs.

The basic idea was to to define classes as follows: A class consists of at least one noun that has received the same answers by at least two annotators. Hence a class is one particular patterned picked from the 720 possible ones. Class 235 – to give an illustration – has the pattern *no, not applicable, yes, not equivalent, yes, no.* This led to 15 classes containing a total of **9,575 distinct** agreeing instances.

In addition to these 9,575 agreeing instances, more than 1,000 noun senses were annotated by all annotators in the group training stage. So, we have 1,000 additional agreeing instances. But, how many classes occur among the 1,000 noun-sense pairs? In the group stage, we can identify 34 different classes, out of which 22 have less than five members. We have decided not to consider classes that occur only in the group stage and have less than five members. So, we will add noun-sense pairs from the group phase, which belong to one of the 15 extracted classes and those classes that have more than five members. The following table summarized the resulting classes, where Y *(yes)*, Num *(Number)*, and Eq *(equivalent)* are short hand for affirmative answers, N, ~Num, and ~Eq are shorthand for negative answers, and NA stands for *not applicable.*

Table 1: Distribution of nouns into classes

Class name	Pattern	four annotators	two annotators (FW, MD)	two annotators (LS, MJ)	group stage	Sum
235	N,NA,Y,~Eq,Y,N	437	3523	3391	674	8025
528	Y,~Num,NA,NA,N,Y	78	705	892	191	1866
510	Y,~Num,Y,Eq,N,Y	14	71	144	85	314
523	N,NA,NA,NA,N,Y	1	68	58	19	146
726	Y,~Num,Y,~Eq,Y,Y	0	48	68	46	162
531	Y,Num,NA,NA,N,Y	1	1	4	6	12
37	N,NA,NA,NA,N,N	1	30	17	11	59
199	N,NA,NA,NA,Y,N	0	2	6	1	9
28	N, NA, N, NA, N, N	0	2	1	1	4
513	Y, Num, Y, Eq, N, Y	0	1	0	0	1
190	N, NA, N, NA, Y, N	0	0	3	8	11
729	Y, Num, Y, ~Eq, Y, Y	0	0	3	0	3
73	N, NA, Y, ~Eq, N, N	0	0	2	1	3
721	N, NA, Y, ~Eq, Y, Y	0	0	1	6	7
353	NA, NA, N, NA, NA, NA	0	0	2	2	4
514	N,NA,N,NA,N,Y	0	0	0	9	9
519	Y,~Num,N,NA,N,Y	0	0	0	25	25
371	NA,NA,Y,NA,NA,NA	0	0	0	7	7
						10,667

As for the class names, the names are artifacts from the analysis of the annotations in R (http://cran.r-project.org/). The numbers might be cryptic at first, but we think that they serve a good purpose in that they allow to define classes without being forced to provide "sensible" names for the classes. In fact, class 235 could be called *fully countable,* but such a simple explanation cannot be provided for each class. The user of the BECL may thus decide for himself or herself, to what kind of classical countability class the nouns in one of the classes should be assigned.

Here are, however, some indications for possible interpretations:

If the answer to test pattern I.1 – as provided in (1) – is *yes* (i.e. Y in the table), then this sense is likely to be a "mass" sense – although there are some caveats to be made about this. If the answer is *no* or *not applicable* then it is likely not to be a "mass" sense.

If the answer to test pattern II.1 (“Can the noun-sense be pluralized?”) is *yes,* then this sense is likely to be a “count” sense – again with some caveats. If the answer to this is *not applicable* then it is likely not to be a “mass” sense.

In a very preliminary analysis, and ignoring all possible caveats, it seems that (at least some of) the classes provided in the BECL can be grouped together as follows:

1. *regular count senses:* classes 235, 721, and possibly the irregular class 371
2. *regular mass senses:* classes 519, 528, 531
3. *senses that are both mass and count:* classes 510 and 726
4. *senses that are neither mass nor count:* classes 37, 190, 199, 514, 523

While we distribute the BECL, the analysis of the data is continued, so future distributions will contain further analyses.

3. What is contained in the BECL?

3.1 Contents of the distribution

The current distribution of BECL (BECL 2.0.zip) consists of the following contents:

- This document.
- The MS Excel (xlsx-format) file classes (15.10.14).xlsx. This spreadsheet contains 18 sheets with the individual classes. For the columns (i.e. features), cf. section 3.2.
- The folder Classes_csv, which contains the 18 classes that we are currently working with in csv format for further processing in R, WEKA, and other relevant programs. Files differ in their size, as can be seen from **Table 1.** The columns are identical to the columns of the sheets in classes (15.10.14).xlsx, except for the added last column, which provides the class name for each entry of the class. This might look redundant. The feature is set here, because the files contained in this folder form the basis for the files master.xlsx, master.csv, and multiples.csv. For the columns (i.e. features), cf. section 3.2.
- The files master.xslx and master.csv (which only differ in their formats, the csv format for use in other programs). Both files contain all nouns in alphabetical order with all columns, as present in classes (15.10.14).xlsx, and the csv-files in Classes_csv. These files can be used to extract nouns that show more than one sense, and a distribution of the senses across different countability classes. For the columns (i.e. features), cf. section 3.2.
- The file multiples.csv contains all nouns that have more than one sense annotated and show a distribution across the countability classes. For the columns (i.e. features), cf. section 3.2.
- LREC_2014.pdf. This is Kiss, Pelletier, and Stadtfeld (2014), which describes the test patterns and provides further information about the project. Information about the project can also be gathered from the following address:
  http://www.linguistics.ruhr-uni-bochum.de/countability/index_en.shtml.

3.2 The structure of the tables

The following structure applies to the following tables:

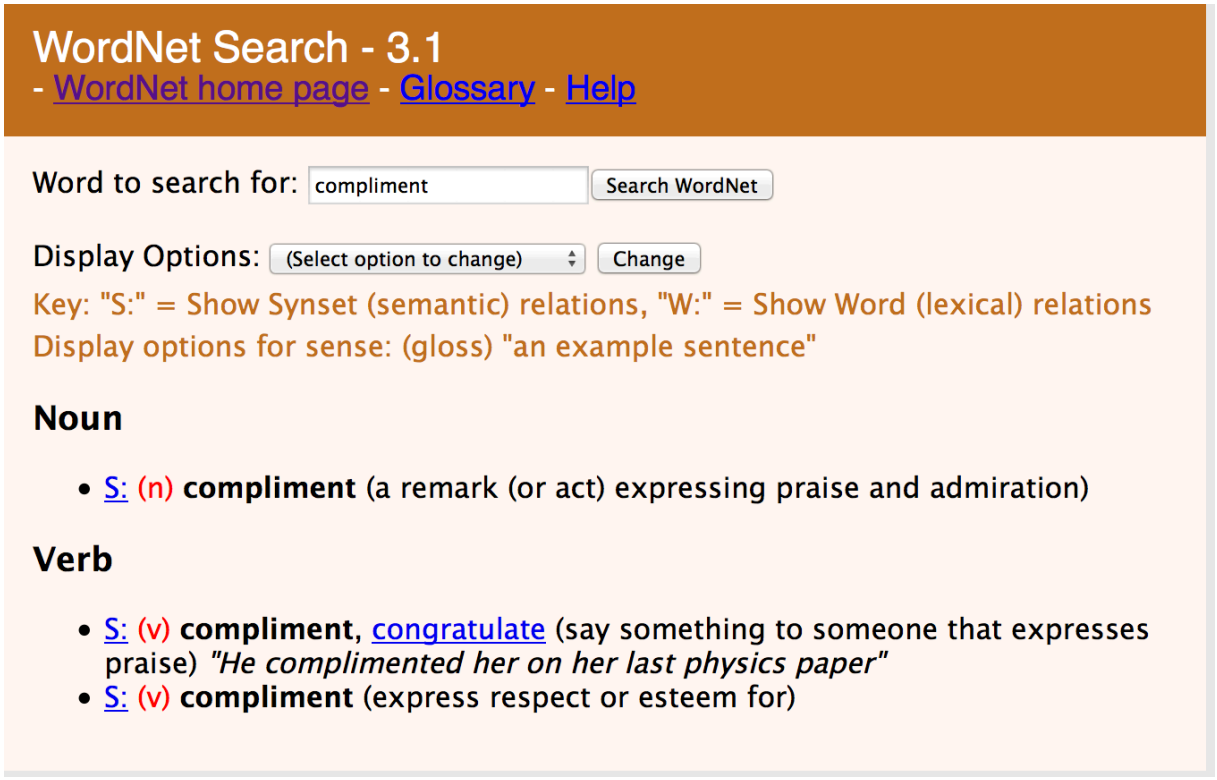
- classes (15.10.14).xlsx

- master.xslx
- master.csv
- All files in Classes_csv

The structure of multiples.csv is described in section 3.3.

ID_AND_SENSE contains a combination of an – arbitrary – ID of the noun and the noun's sense. The nouns have been extracted from the OpenANC according to the following rules: The noun should occur not less than 10 times, and at least one of its senses should be given in WordNet. The total number of lemmata satisfying these conditions is approximately 42,000. From these 42,000 lemmata, BECL contains 6,650 lemmata with up to four senses.

The sense refers to the sense in WordNet; note that senses in WordNet are not numbered, and that we only consider senses for nouns. Consider the following screenshot from WordNet.



Here, *compliment* is given with three senses, but only one of them is for a noun. Hence, compliment is assumed to have only one sense in BECL.

noun provides the lemma of the noun

wordnet_senseindex_number indicates which of the senses of WordNet is indicated. Please note that senses are not numbered in WordNet, so we have just repeated the order provided in a WordNet search, as illustrated for *elimination* in the following screenshot:

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: elimination Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **elimination**, riddance (the act of removing or getting rid of something)
- S: (n) **elimination**, evacuation, excretion, excreting, voiding (the bodily process of discharging waste matter)
- S: (n) **elimination**, reasoning by elimination (analysis of a problem into alternative possibilities followed by the systematic rejection of unacceptable alternatives)
- S: (n) **elimination** (the act of removing an unknown mathematical quantity by combining equations)
- S: (n) **elimination**, liquidation (the murder of a competitor)

Please note that *elimination* is given with five senses in WordNet, but we have only considered the first four. The numbering of the senses in BECL reflects the order of presentation in WordNet search.

wordnet_description provides the pertinent sense of the noun-sense pair. In the actual annotation, the annotators were provided with a noun and one of its senses in an almost arbitrary fashion.

occurrences_in_oanc_total provides information about the total frequency of the noun lemma in Open ANC.

occurrences_singular_in_oanc provides information about the frequency of singular – or unmarked in the case of possible mass nouns – occurrences of the noun lemma in Open ANC.

occurrences_plural_in_oanc provides information about the frequency of plural occurrences of the noun lemma in Open ANC.

The following six columns provide the actual results of the annotation. The columns indicate the respective answered to the six test patterns, as described in Kiss, Pelletier, and Stadtfeld 2014. The annotators provided under **annotators** have carried out the annotation individually**,** but all annotators listed have agreed on the values, so there was no necessity to list the values for the individual annotators.[1]

Test_I_1 The answer to the first question of the first test pattern.

Test_I_2 The answer to the second question of the first test pattern.

Test_II_1 The answer to the first question of the second test pattern.

[1] One of the reasons that not all 15,000 noun-sense pairs that have been annotated are contained here is simply that the annotators did not agree on their annotations.

Test_II_2 The answer to the second question of the second test pattern.

Test_III_1 The answer to the first question of the third test pattern.

Test_III_2 The answer to the second question of the third test pattern.

The following six columns are contained but not currently employed in the analysis. The column **comment** can be used for comments by the annotators. The column **idiomatic** was used to indicate possible idiomatic senses. The columns **nominalization, result_state, process,** and **act_event** have been used to approximate an analysis of nominalizations, but the analysis is superficial at best.

The column **Phase_No** provides information about the annotation phase, and allows to derive the annotators. In the present distribution, the annotators are listed in the column **annotators,** so that the column **Phase_No** is possibly superfluous.

The column **class** is not contained in the file classes (15.10.14).xlsx. It contains the class number that has been arbitrarily assigned by R to the combination of the six test pattern answers.

3.3 The table multiples.csv

The table **multiples.csv** contains all noun lemmata that show more than one sense, where at least one of the senses is annotated differently from other senses. The classes 510 and 726 contain nouns that we call *dual-sense nouns,* i.e. nouns which show one particular sense that allows both a count and a mass interpretation. An example for a dual-sense noun in class 726 is *cable* with the sense definition "a conductor for transmitting electrical or optical signals or electric power" (cf. also Rothstein 2010). The nouns in multiples.csv contain **lemmata** that show different properties with respect to count and mass if their senses are discerned, they can thus be called *dual-use* nouns. The table can be used to identify such lemmata, and the lemmata can be searched in master.xlsx or master.csv. In R, the nouns in multiples.csv can be used to subset the relevant data from master csv.

The table **multiples.csv** contains the following information:

noun: the noun lemma

c235-c513: The countability class in which one of the senses occurs.

sum: the number of senses annotated.

max: the maximal number of senses showing up in one of the classes.

References

Alan Bale and David Barner. 2009. The interpretation of functional heads: Using comparatives to explore the mass/count distinction. *Journal of Semantics*, 26:217– 252.

David Barner and Jesse Snedeker. 2005. Quantity judg- ments and individuation: Evidence that mass nouns count. *Cognition*, 97:41–66.

Kiss, Pelletier, and Stadtfeld (2014)

Susan Rothstein. 2010. Counting and the mass-count dis- tinction. *Journal of Semantics*, 27:343–397.