

A Sense-based Lexicon of Count and Mass Expressions: The Bochum English Countability Lexicon

Tibor Kiss*, F. Jeffry Pelletier†, Halima Husic*, Johanna M. Poppek*, R. Nino Simunic*

* Ruhr-Universitaet Bochum, Germany

† Simon Fraser University Vancouver, Canada

{tibor, husic, poppek, simunic}@linguistics.rub.de, jeffpell@sfu.ca

Abstract

The present paper describes the current release of the *Bochum English Countability Lexicon* (BECL 2.1), a large empirical database consisting of lemmata from Open ANC (<http://www.anc.org>) with added senses from WordNet (Fellbaum, 1998). BECL 2.1 contains ≈ 11,800 annotated noun-sense pairs, divided into four major countability classes and 18 fine-grained subclasses. In the current version, BECL also provides information on nouns whose senses occur in more than one class allowing a closer look on polysemy and homonymy with regard to countability. Further included are sets of similar senses using the Leacock and Chodorow (LCH) score for semantic similarity (Leacock & Chodorow, 1998), information on orthographic variation, on the completeness of all WordNet senses in the database and an annotated representation of different types of proper names. The further development of BECL will investigate the different countability classes of proper names and the general relation between semantic similarity and countability as well as recurring syntactic patterns for noun-sense pairs. Our current work on those patterns concerning mass nouns is briefly discussed pointing to further research. The BECL 2.1 database is also publicly available via <http://count-and-mass.org>.

Keywords: countability, mass/count distinction, questionnaire-based lexical resource, lexicon, semantic similarity, proper names

1. Introduction

The Bochum English Countability Lexicon (BECL) provides a large-scale publicly available resource for the analysis of the count/mass distinction (cf. <http://count-and-mass.org>). BECL addresses three problematic aspects in assessing the count/mass distinction: First, that the distinction can be applied to the lemma or lexeme (as e.g., in Grishman et al. (1994) and Bond (2007)). Secondly, the popular view that the distinction is a binary one (as exemplified in Borer, 2005 and ISO Space, 2014), and thirdly, that it can be analysed by looking at a small sample of staple nouns only.

BECL replaces this methodology with over 11,800 annotated noun-sense pairs. A noun-sense pair is a combination of a noun with an individual sense of that noun. Noun-sense pairs in BECL fall into four major countability classes that are subdivided into 18 minor countability classes. The original annotations are provided with the resource, thus allowing users to assess the data directly. Further information on the initial annotation steps including inter-annotator agreement can be found in Kiss, Pelletier, and Stadtfeld (2014). An illustration of BECL is provided in (1).

This paper describes the current release of BECL, including the definition of noun-sense pairs, the characterization of classes, nouns whose senses occur in more than one class, the usage of the Leacock and Chodorow (LCH) score for semantic similarity (Leacock and Chodorow, 1998), and the representation of proper names.

2. Noun Senses and Countability Classes in BECL

All annotation in BECL applies at the level of the noun-sense pair. In replacing lexemes or lemmas as the locus of

the count/mass distinction by noun-sense pairs, we are able to represent polysemy that has an effect on the count/mass distinction. Annotators were not only provided with a lexeme, but also with the relevant sense derived from WordNet (Fellbaum, 1998; <http://wordnet.princeton.edu>).

BECL provides a four-way distinction into major classes: noun-senses that are count, noun-senses that are mass, noun-senses that are both count and mass, and noun-senses that are neither count nor mass (cf. (2)). Noun-sense pairs are represented through $noun_{number}$, where *noun* is the form, and *number* is the sense number provided in WordNet.

(2)

- a. count: *academy*₂ (an institution for the advancement of art or science or literature)
- b. mass: *fatigue*₁ (temporary loss of strength and energy resulting from hard physical or mental work)
- c. both: *aerosol*₁ (a cloud of solid or liquid particles in a gas)
- d. neither: *country*₄ (an area outside of cities and towns, e.g., *John went for a walk in the country.*)

The usefulness of noun-sense pairs becomes apparent when cases of polysemy or homonymy are considered (cf. (3)). The two interpretations of *matter* fall into different countability classes. But without a distinction at the sense level, the multiple assignment of the form, given different senses, would not be discernible from nouns where the very same sense behaves like count *and* mass - as illustrated by *aerosol*₁ in (2c).

(3)

- a. *matter*₁ (count: *a vaguely specified concern; "several matters to attend to"*)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	W	Z	AB	AC	AD	AE
1	id_and_sense	lemma	sense_id	sense_desc	sense_desc	sense_desc	similar(CF>2.5)	occurrences_in_sense	occurrences_in_sense	occurrences_singular_in_sense	Syn1	Sem1	Syn2	Sem2	Syn3	Syn4	proper name	class	complete	verification_in_writing		
2	25085.3	aa	3	an associate degree in arts	associate_in_arts.n.01	academic_degree.n.01, associate_degree.n.01, associate	306	305	1	no	not appl	yes	not equiv	yes	no			235				
3	40178.1	abbreviation	1	word or phrase a shortened form of a state or condition markedly different from the norm	abbreviation.n.01	word.n.01, form.n.01, plural.n.01, singular.n.01, ghost_w	140	33	107	no	not appl	yes	not equiv	yes	no			235				
4	20030.1	aberrancy	1	a state or condition markedly different from the norm	aberrance.n.01	acardia.n.01, acephalia.n.01, acorea.n.01, acromicria.n.01	14	3	11	no	not appl	yes	not equiv	yes	no			235	Yes			
5	24831.1	aberration	1	a disorder in one's mental state	aberrance.n.01	acardia.n.01, acephalia.n.01, acorea.n.01, acromicria.n.01	30	7	23	no	not appl	yes	not equiv	yes	no			235	Yes			
6	24831.2	aberration	2	an optical phenomenon resulting from the failure of a lens or mirror to produce a good image	aberration.n.02	disorder.n.01, mental_disorder.n.01, aspergers_syndrome	30	7	23	no	not appl	yes	not equiv	yes	no			235	Yes			
7	24831.3	aberration	3	an abnormal physical condition resulting from defective genes or developmental deficiencies	aberration.n.03	physical_phenomenon.n.01, aberration.n.03, absorption	30	7	23	no	not appl	yes	not equiv	yes	no			235	Yes			
8	42805.1	abnormality	1	a reformer who favors abolishing slavery	abnormality.n.01	wakefulness.n.01, condition.n.01, bloodiness.n.01, incom	288	62	226	no	not appl	yes	not equiv	yes	no			235				
9	14926.1	abolitionist	1	failure of a pianist of pivoting 360 degrees, especially in a military formation	abolitionist.n.01	disruptant.n.01, abolitionist.n.01, birth_control_campaign	21	11	10	no	not appl	yes	not equiv	yes	no			235	Yes			
10	30579.2	abortion	2	a major change in attitude or principle or	miscarriage.n.01	happening.n.01, failure.n.02, downfall.n.01, flame-out.n.c	374	470	104	no	not appl	yes	not equiv	yes	no			235				
11	21433.1	about-face	1		about-face.n.01	change_of_direction.n.01, reversion.n.04, about-face.n.01	13	13	0	no	not appl	yes	not equiv	yes	no			235	Yes			

(1) User Interface to BECL

b. matter₃ (mass: *that which has mass and occupies space*; “physicists study both the nature of matter and the forces which govern it”)

Nouns that can be both count and mass are often referred to as *dual life nouns*. In the literature, this distinction is always drawn on the basis of the noun-form only, so it always remains unclear whether the noun with a given sense is both a count and a mass noun or whether its assignment to the classes depends on the fact that it can be interpreted with different senses. In BECL, a distinction is drawn between *dual life nouns* proper, which are noun-sense pairs that can be both count and mass, and (what we call) *multiples* where a form *noun* combines with two different senses to yield noun-sense pairs in two different countability classes.

The fine-grained assignment to classes in BECL is based on the annotation described in Kiss, Pelletier, and Stadtfeld (2014). Provided with noun-sense pairs, the annotators did not answer questions about countability directly, but had to insert noun-sense pairs into syntactic patterns, and had to explore certain semantic consequences of the insertion. The syntactic patterns are:

- **Syntactic Pattern 1 (Syn1):**
Can the noun-sense pair in its singular form appear together with *more*?
- **Syntactic Pattern 2 (Syn2):**
Can the noun-sense pair in its plural form appear together with *more*?
- **Syntactic Pattern 3 (Syn3):**
Can the noun-sense pair in its singular form and combined with an indefinite determiner be the syntactic subject of a definition or characterization?
- **Syntactic Pattern 4 (Syn4):**
Can the noun-sense pair in its singular form but without a determiner be the syntactic subject of a definition or characterization?

The answers to Syn1 and Syn2 invite further semantic questions that are represented in Sem1 and Sem2: If

the answer to Syn1 was *yes*, the annotator had to decide whether the comparison was based on number or another mode of measurement (**Sem1**). If the answer to Syn2 was *yes*, the annotator had to decide whether the sentence is semantically equivalent to a sentence with an explicit classifier (**Sem2**). For the noun-sense pairs in (2) and (3), the following annotations emerge (where NA stands for *not applicable*)¹.

Noun	Sense	Syn1	Sem1	Syn2	Sem2	Syn3	Syn4	Class
<i>academy</i>	2	no	NA	yes	-eq	yes	no	235
<i>fatigue</i>	1	yes	-num	NA	NA	no	yes	528
<i>aerosol</i>	1	yes	-num	yes	eq	no	yes	510
<i>country</i>	4	no	NA	NA	NA	no	yes	523
<i>matter</i>	1	no	NA	yes	-eq	yes	no	235
<i>matter</i>	3	yes	-num	NA	NA	no	yes	528

(4) Countability Classes in BECL

Classes correspond to unanimous answers for the six patterns. From logical set of 80 classes, 18 classes can be defined on the basis of unanimous annotations by at least two annotators (cf. Kiss, Pelletier, and Stadtfeld, 2014). The resulting 18 fine-grained classes are grouped into the four major classes already illustrated in (2). The distribution of the four major classes among all annotated noun-sense pairs is provided in (5).

All count senses have a plural (Syn2 = *yes*). All mass senses can be combined with *more* in the singular (Syn1 = *yes*), have to be used without an indefinite determiner (Syn3 = *no*), and can be used without a determiner at all (Syn4 = *yes*). Senses that are both mass and count occur in the singular and plural with *more* (Syn1 = *yes* and Syn2 = *yes*), and they also may appear without a determiner at all (Syn4 = *yes*). Senses that are neither mass nor count must not appear with *more* in the singular (Syn1 = *no*, against *mass*), and do not form a plural (Syn2 = *no*, against *count*).

¹The class names (235, 538 etc.) are an artifact of the initial extraction process carried out in R (<https://cran.r-project.org>).

Class	Frequency	Subclasses
regular count	8,434	4
regular mass	2,427	3
both count and mass	699	4
neither count nor mass	315	7
Sum	11,875	18

(5) Distribution of Countability Classes

3. What is Contained in BECL?

BECL is released as a set of tables in csv-format with tables for each fine-grained class, a table containing all noun-sense pairs, and a table consisting of nouns that show senses that appear in different classes (as illustrated in (3)). The tables contain the following information:

- The noun form, its sense number in WordNet, and its sense description in WordNet, as illustrated above in (2).
- The total number of senses in WordNet and the synset for the sense
- The set of similar synsets measured according to the LCH score (cf. section 4).
- An indicator of whether all senses of a noun listed in WordNet have been annotated.
- The frequency distribution of the lemma, the singular and the plural form in OpenANC (www.anc.org).
- The outcome of the tests Syn1, Syn2, Syn3, Syn4, Sem1 and Sem2.
- Whether the noun in question is likely have a proper name as a homograph (cf. section 5). Users can select whether they want to include proper names.
- Information about the annotation process.
- An indicator of whether the noun occurs in more than one orthographic form (and is thus a duplicate apart from its orthography).

4. Mining for Similarities

BECL as a sense-based lexicon offers a distinctive advantage over traditional methods to assess the count/mass distinction in that interesting formal and semantic similarities can be extracted from the data. These similarities would presumably not have been detected without such a systematic collection of data.

Similar senses for each noun sense were extracted via the NLTK WordNet Module (Bird et al., 2009) using the LCH Similarity Measure (Leacock and Chodorow, 1998). LCH compares two senses based on the shortest path (nodes) between them and the maximum depth of the taxonomy. The resulting sets with a comparatively high threshold of 2.5 received form relatively homogenous clusters or semantic groups.

We have already discussed so-called *dual life nouns* in (2c) and polysemous nouns in (3). A dual life noun shows a pair of a unique form and a unique sense but is both count and mass. A polysemous noun shows one form that is related to a set of senses so that the senses belong to different countability classes.

Since BECL provides information about semantic similarity in terms of the LCH score, BECL can also be used to identify a sense $sense_1$ that is distributed among different forms $\langle form_1, \dots, form_n \rangle$ so that the resulting groups $\langle \langle form_1, sense_1, \dots, \langle form_n, sense_n \rangle \rangle$ occur in different countability classes. Examples are provided in (6).

(6)

- a. $\langle poison_1 \rangle$ 510 (both) vs. $\langle toxicant_1 \rangle$ 235 (count)
- b. $\langle sensibility_3 \rangle$ 528 (mass) vs. $\langle sensitivity_1 \rangle$ 510 (both)
- c. $\langle disruption_1 \rangle$ 726 (both) vs. $\langle interruption_1 \rangle$ 235 (count)
- d. $\langle pay_1 \rangle$ 528 (mass) vs. $\langle salary_1 \rangle$ 235 (count) and $\langle wage_1 \rangle$ 235 (count)

BECL thus provides a basis for further investigations identifying why identical meanings (in the sense of identical synsets) lead to different countability classes.

5. Proper Names

The nouns for BECL were initially compiled with the proviso that the nouns occur more than 10 times in OpenANC, and that a sense definition is provided in WordNet. Hence, BECL also contains proper names. Since proper names are not identified as such in WordNet, BECL has been annotated independently by two annotators for proper names, organisations, and locations, with an inter-annotator agreement of 0.82 (Krippendorff's α , Artstein and Poesio, 2008). Current research assumes implicitly that proper names do not fall under the count/mass distinction since they carry a uniqueness implication, which prohibits counting their referents even though they are bounded entities. One would thus expect that proper names fall into the major class *neither count nor mass*, which is correct for the majority of proper names identified in BECL. However, BECL also contains proper names that are members of the other classes, as illustrated in (7).

(7)

- a. $voodoo_3$: a religious cult practiced chiefly in Caribbean countries (especially Haiti); involves witchcraft and animistic deities 528 (mass)
- b. $coke_2$: Coca Cola is a trademarked cola 510 (both)
- c. $exec_1$: the chief executive department of the United States government 235 (count)

BECL provides a basis to investigate why certain proper names fall under the jurisdiction of the count/mass distinction.

6. Future Directions: Parsing, Grinding, and Sorting

For future releases, we will add syntactic information about recurring patterns for noun-sense pairs that fall into unique countability classes; this should provide some indication how likely it is for noun to take part in so-called grinding or other type coercions (cf. Cheng et al., 2008, Djalali et al., 2011). Since we expected a much greater variety on data concerning mass-to-count than count-to-mass shifting, we started our analysis with corpus examples regarding lemmata from countability subclass 528 (regular mass) on the parsed data.

We focused on nouns for which all senses present in WordNet have been annotated uniquely as belonging to class 528 (so the nouns would show the features *complete* == *yes* and *multiple* == *NA*). Against this background, deviant behavior is characterized by either of the following properties:

- The noun occurs with an indefinite determiner or
- the noun occurs in the plural.

We have parsed the OANC corpus data using the transition-based, neural network powered parser (Chen & Manning, 2014) developed by the Stanford NLP Group (<http://nlp.stanford.edu>). The parser uses a (standard²) language model trained on the Penn Treebank and the original linguistic annotation contained in the OANC to produce dependency parses using CoNLL Dependencies of our target sentences. For the relevant nouns in class 528, OANC provided 100,193 sentences for 810 lemmata in total.

After removal of false positives from the data (e.g., unidentified compound nouns, false POS-Tags, suspicious classification of 528 lemmata), a total of 374 unique lemmata of class 528 occurring in 5,536 sentences were left that appear in conspicuous syntactic patterns indicating a mass-to-count shift for the respective lemma.

Given current observations, the large majority of these shifting examples can be divided into three main categories:

1. A mass noun is shifted into a count noun because of an arising interpretation as a unit or “specific quantity of a substance” (as in examples in (8)) (cf. De Belder, 2008; Borer, 2005; Payne & Huddleston, 2002).

(8)

- a. For the *Mysidopsis bahia* and the *Cyprinodon variegatus* test methods, the effluent sample was a municipal *wastewater* spiked with KCl. (...)
- b. Homeostatic regulation of sleep can be quantified objectively after a period of enforced *wakefulness*. (...)
- c. So there is a double *nostalgia* involved in reading these books – nostalgia for one’s own childhood and nostalgia for the timeless realm of classic children’s fiction.

²Included in the parser software package (<http://nlp.stanford.edu/software/mndep.shtml>)

Although this interpretation is mostly referred to as appearing with distinct measure phrases, *unit*-readings, as seen in (8), do not seem to require one. Since this phenomenon is widely discussed in a number of publications (cf. above), it is nonetheless notable that the examples from the 528 subclass only contained very few cases that strongly suggest a *unit*-interpretation (< 50) after the filtering. Reasons could be that without a specific measure phrase or in case of abstract nouns, they are easily confused with examples of the sort discussed in (9) or (10).

2. The noun acquires a *kind-of*-reading by defining a certain element of a greater variety that is represented by the bare mass noun. (cf. De Belder, 2008; Borer, 2005; Payne & Huddleston, 2002).

(9)

- a. The ability to consider the effect of time-variations on the robustness of the system is one great advantage of the SSV over other *methodologies*.
- b. In contrast to other Italian cities weakened by internal *rivalries* and unstable government, Rome drew strength from a solid aristocracy of consuls and senate ruling over plebeians proud of their Roman citizenship and only rarely rebellious.
- c. The universe, in short, is breaking *symmetries* all the time by generating such novelties, creating distinctive molecules or other forms which had never existed before.

As mentioned above, *kind-of*-readings cannot be in all cases easily distinguished from other shifting types. However, roughly 25% of the final examples showing mass-to-count shifting clearly allow this interpretation.

3. The shifting is based on a further interpretations as an *act*, *event* or *result* relating to its abstract content as a bare mass noun (cf. Payne & Huddleston, 2002 on “event instantiation” of mass nouns):

(10)

- a. The reaction products were purified by means of three repeated gel *chromatographies* using water saturated Sephadex G-50 in Millipore/ Multiscreen filtration plates according to the instructions provided by the supplier and dried under vacuum.
- b. Not for him the *excesses* of Gothic supernaturalism or ridiculous stories of ghosts, stick figures, and things that go bump in the night.
- c. Achieving consistent *measurements* presents a major problem because two trained fitters will generally come up with important differences in the body measurements of the same person.

This type of mass-to-count shifting is rarely discussed in the literature, and when it is, it is usually described as a restricted extension to certain categories of nouns,

and as neither regular nor predictable (e.g., by Payne & Huddleston, 2002). Nonetheless, about 15% of our final data can be interpreted in terms of *instantiation* of a mass noun.

Based on these rough observations regarding shifted interpretations for mass nouns, further work on the issue will include research on the relation between shifting patterns and semantic similarity as well as the respective nominal category, especially concerning the case of event-instantiation.

7. Next Steps

With regard to the theoretical work underlying the development of BECL, the next steps will include the further investigation of the different countability classes of proper names, and the general relation between semantic similarity and countability. Special attention to the notion of “abstract nouns” will also feature in our investigations. Also, the analysis of syntactic patterns and shifting between mass and count will be extended to lemmata from other countability subclasses.

8. Bibliographical References

- Artstein, Ron; Poesio, Massimo (2008) Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, pp. 555–596.
- Bird, Steven; Klein, Ewan; Loper, Edward (2009) *Natural Language Processing with Python*. Sebastopol, CA: O’Reilly Media Inc.
- Bond, Francis X. (2007) *Translating the Untranslatable: A Solution to the Problem of Generating English Determiners*. Stanford: CSLI Press.
- Borer, Hagit (2005) *Structuring Sense. Vol. I: In Name Only*. Oxford: Oxford University Press.
- Chen, Danqi; Manning, Christopher D. (2014) A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP 2014*, <http://cs.stanford.edu/danqi/papers/emnlp2014.pdf>.
- Cheng, Lisa, Doetjes, Jenny; Sybesma, Rint (2008) How universal is the universal grinder? *Linguistics in the Netherlands 2008*, pp. 50–62, Amsterdam: John Benjamins.
- De Belder, Marijke (2008) Size matters: Towards a syntactic decomposition of countability. Abner, Natasha & Jason Bishop (eds.) *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Somerville: Cascadilla Proceedings Project.
- Djalali, Alex; Grimm, Scott; Clausen, David; Levin, Beth (2011) What can be ground? Noun type constructions, and the universal grinder. *Proceedings of the Berkeley Linguistics Society*. University of California, Berkeley.
- Fellbaum, Christiane (ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Grishman, Ralph; Macleod, Catherine; Meyers, Adam (1994) *COMLEX Syntax: Building a Computational Lexicon*. Coling 1994, Kyoto.
- Huddleston, Rodney; Pullum, Geoffrey K. (eds.) (2002) *Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- ISO 2014. ISO 24617-7:2014 *Language resource management - Part 7: Spatial information (ISOspace)*. ISO, Geneva. ISO Working Group: TC 37/SC 4/WG 2.
- Kiss, Tibor; Pelletier, F. Jeffry; Stadtfeld, Tobias (2014) Building a Reference Lexicon for Countability in English. In: *Proceedings of LREC 2014*. Reykjavik, Iceland.
- Leacock, Claudia; Chodorow. Martin (1998) Combining Local Context with WordNet Similarity for Word Sense Identification. Fellbaum, C. (1998), pp. 265–283.
- Payne, J.; Huddleston, Rodney (2002). Nouns and noun phrases. Huddleston, R.; Pullum, G., pp. 323–524.