Journal of English for Academic Purposes 14 (2014) 118–123

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Forum

# A closer look at laughter in academic talk: A reader response

Ray Carey*

*Department of Modern Languages, P.O. Box 24, Room C520, 00014 University of Helsinki, Finland*

ARTICLE INFO

*Article history:*
Received 31 December 2013
Received in revised form 7 March 2014
Accepted 14 March 2014

*Keywords:*
Laughter frequency
Academic discourse
British Academic Spoken English (BASE) corpus
Michigan Corpus of Academic Spoken English (MICASE)
Corpus of English as a Lingua Franca in Academic Settings (ELFA)

ABSTRACT

In this reader response, the quantitative findings from Nesi (2012) on laughter in academic talk are reproduced and reconsidered. An anomalous word count cited for the BASE lecture subcorpus is corrected and the normalized frequency of laughter in these lectures is shown to be twice than what is reported. When the per-minute frequency of laughter is also corrected, the frequencies of laughter in BASE and MICASE lectures appear to be nearly identical, contrary to Nesi's (2012) claims. Additionally, her suggestion that laughter may be less frequent in English-medium lectures outside of L1 English settings is examined in light of data from the ELFA corpus of academic speech events recorded in Finland. While the monologic lectures in ELFA show a lower frequency of laughter than the monologic lectures in BASE and MICASE, a strong preference for laughter episodes is found in the dialogic ELFA lecture discussions. The high frequency of laughter in ELFA discussions is especially evident in the seminar files, where laughter frequencies eclipse those found from seminars in the native-speaker corpora.



## 1. Introduction

In the June 2012 edition of the *Journal of English for Academic Purposes* (vol. 11/2, 79–89), Hilary Nesi presented her findings on laughter in university lectures, drawing primarily from data in the corpus of British Academic Spoken English (BASE). Her findings include a quantitative analysis of the frequency of laughter in these speech events as well as a qualitative analysis of the types and functions of these laughter episodes. Apparent inconsistencies in the quantitative findings motivated me to reproduce her study, the results of which I have earlier discussed on the ELFA project research blog (http://elfaproject.wordpress.com/2013/07/27/laughter-in-academic-talk-brits-yanks-elf-compared/). In this reader response, I report these reproduced quantitative findings and further consider Nesi's (2012: 87) claims about the frequency of laughter in academic English speech events outside of L1 English settings. For this discussion, I introduce findings from the ELFA (2008) corpus concerning frequencies of laughter from academic discourse in which English is used as a lingua franca (ELF).

The BASE corpus is reported to have an official word count of 1 644 942 words (Nesi & Thompson, 2006), a figure also cited by Nesi (2012: 81). However, in her quantitative analysis of laughter frequencies in the BASE lecture subcorpus, a total word count of 2 646 920 words is given for this BASE lecture subcorpus alone, a million words higher than the total BASE corpus word count (see Tables 2 and 3 in Nesi, 2012: 83). This inconsistency is not accounted for elsewhere in the article, and this apparently erroneous word count is used to calculate similarly erroneous normalized frequencies of laughter, which should thus be artificially low. Taking advantage of the free availability of BASE, I downloaded the XML version of the corpus, which

* Tel.: +358 44 923 8802 (mobile).
*E-mail address:* ray.carey@helsinki.fi.

http://dx.doi.org/10.1016/j.jeap.2014.03.001
1475-1585/

Nesi also used for her study, to reproduce her quantitative findings. These findings and the methodology by which I reproduced the study are discussed in Section 2, followed in Section 3 by my original findings on laughter in the ELFA corpus.

## 2. Corrected quantitative findings & methodological considerations

In order to calculate the frequencies of various types of laughter in BASE lectures, Nesi used the WordSmith Tools software to search for the XML tags encoding laughter. Corpora encoded in XML format are readily used in concordance software by selecting an option to ignore all metadata within so-called "angle brackets," i.e. the less-than (<) and greater-than (>) symbols. However, unlike other plain text corpus formats, XML brings added value as structured, well-formed, computer-readable data. Searching for XML tags as strings of text in a concordancer misses out on this value. In addition to the common designator of "tags", XML elements are best understood as programming objects – nodes within a structured tree that can be directly accessed by means of a programming language. This approach avoids potential problems with variations in the XML markup and opens new analytical possibilities.

Additionally, a potential problem with treating XML tags as strings of text lies in obtaining total word counts. If a concordancer is used to generate a word list without ignoring the contents within angle brackets, the XML will also be tokenized at whitespaces and result in an inflated token count. This was my first hypothesis of how Nesi obtained a lecture word count of 2 646 920 words. However, a word list in WordSmith Tools derived from the BASE lecture subcorpus with XML tags included yields a count of 1 816 904 tokens used for the word list, with 2 231 912 running words. Even this does not approach Nesi's reported word count, and I have been unable to find an explanation that would account for her findings.

For my own reproduction of Nesi's quantitative findings, I wrote a program in the programming language Python to parse the BASE XML corpus with the lxml library (http://lxml.de). This Python code is available from the ELFA project blog post referenced above. My approach to tokenizing the BASE XML files was to 1) exclude the contents of the <teiHeader> element, 2) strip all XML markup from the text, and 3) tokenize what remained at whitespaces. This resulted in a BASE lecture subcorpus token count of 1 208 641 – reasonably close to the official BASE holdings figure of 1 212 251 tokens (spreadsheet available online: http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/holdings/base_corpus_holdings.xls, accessed 28.10.2013). Unable to find further documentation on how the official BASE counts were derived, I have used my own token counts in the figures which follow.

In order to obtain reliable counts of the XML tags designating laughter, the customized Python program iterates through each BASE XML file, identifies the <vocal> and <shift> elements, and checks the attributes on each of these elements for values of "laugh" or "laughter" (see Nesi, 2012: 82 for more description of this markup). These are counted according to the academic domain in which they occur (as in Nesi, 2012), frequencies are normalized, and findings are output to text as comma-separated values. This output reproduces the data in Nesi's Tables 2 and 4 (2012: 83–84), the results of which are shown here in Table 1. Raw counts for tokens and laughter elements in the BASE corpus are subdivided by event type (lecture or seminar) and by the four academic domains in the corpus. Finally, the standardized frequencies are shown as laughter elements per 10 000 tokens, with the standardized frequencies from Nesi's article shown on the far right for comparison.

My count of 2187 laughter elements is four higher than that of Nesi. She has counted an extra laughter tag in the life sciences lecture subcorpus, and I am unable to account for that. My laughter count for BASE seminars is five higher than that of Nesi, but this is due to a single file (sssem006.xml) that is not well-formed XML. Among other markup problems, five laughter elements are annotated in error in this file. I corrected these by hand, as only well-formed XML can be parsed in a programming environment. Oddly, the BASE XML corpus is released with a DTD (Document Type Declaration) file, which can

**Table 1**
Reproduction of the findings from Nesi (2012).

| Domain | Tokens | Total laughs | Laughs/10k tokens | Nesi (2012) per 10k |
|---|---|---|---|---|
| **BASE lecture subcorpus** | | | | |
| Arts/humanities | 308 126 | 306 | 10 | 5 |
| Life sciences | 291 997 | 635[a] | 22 | 10 |
| Physical sciences | 255 256 | 226 | 9 | 4 |
| Social sciences | 353 262 | 833 | 24 | 11 |
| **Total** | **1 208 641** | **2000** | **17** | **8** |
| | | | | |
| **BASE seminar subcorpus** | | | | |
| Arts/humanities | 123 289 | 18 | 1 | 1 |
| Life sciences | 129 153 | 83 | 6 | 6 |
| Physical sciences | 76 716 | 11 | 1 | 1 |
| Social sciences | 95 344 | 75[b] | 8 | 6 |
| **Total** | **424 502** | **187** | **4** | **4** |
| | | | | |
| **BASE totals** | **1 633 143** | **2187** | **13** | |

[a] Nesi's count was 636 for a total of 2001 laughs. I am unable to account for this disparity in our counts.
[b] Nesi's count was 70, but this is because a single file (sssem006.xml) is not well-formed XML. Several laughter elements have been tagged in error. Once corrected manually, the count comes to 75.