Characteristic words for boys and girls in COLT

Koenraad De Smedt

October 4, 2023

Abstract

This paper presents corpus data for 12 words that are assumed to be characteristic for either boys' or girls' speech. From the Bergen Corpus of London Teenage Language (COLT), word occurrences were retrieved, together with the gender of the speakers. The distribution of absolute counts of gender relative to word was obtained. In order to study differences in the use of words between genders, weighted percentages were computed to take into account group sizes.

1 Introduction

It is often assumed that the active vocabulary use of one social group may differ from that of another, as distinguished by age, gender or other parameters. Lexical differences related to gender have been borne out in earlier studies, for instance, a study of telephone conversations (Boulis and Ostendorf, 2005).

The present study has an anecdotal character. It takes as a starting point two sets of words, each one hypothesized to be characteristic for a gender group. The words in (1) are assumed to be characteristic for teenage girls, while those in (2) may be characteristic for boys of the same age group.[1]

(1) cat, clothes, kiss, love, model, phone

(2) beat, bloke, cool, crap, football, music

The aim of this paper is to demonstrate how observations of these 12 words can be obtained from a corpus, analyzed quantitatively and visualized by a Python script.

2 Data

The Bergen Corpus of London Teenage Language (COLT) was collected in 1993 through an initiative led by Anna-Brita Stenström of the University of Bergen. COLT consists of spontaneous conversations between 31 volunteering 13 to 17 year old boys and girls from several school districts in London (Stenström, Andersen, and Hasund, 2002). The speech material of about half a million words was orthographically transcribed by trained transcribers employed by the Longman Group for the British National Corpus (BNC). Every utterance was tagged with gender, age, and other parameters, although in some cases the value of a parameter is unknown. The orthographically transcribed material was subsequently submitted to careful editing in Bergen, where the resulting corpus was initially released on CD-ROM in 1999. Later the corpus has become accessible at CLARINO, where it is searchable online in the Corpuscle tool[2] (Språkkontakt og ungdomsspråk i Norden, 1999). Corpuscle provides several query and analysis modes (Meurer, 2012).

For the present study, the set of words in (1–2) was converted to a query by means of a simple Python program and the query was entered in Corpuscle. Searching in COLT retrieved all occurrences of these words. The distribution of gender relative to word was computed in Corpuscle, with results shown in Figure 1. The gender counts are coded as female (f) or male (m), but some observations do not have a gender value (-), because annotators could not determine this attribute in some cases.

The absolute counts are in black, whereas the percentages in red are distributions of counts within each row. These numbers are however not a perfect basis for comparing words by gender because the total sizes of corpus materials for each gender group are not the same. Searching for all words in the corpus by means of the query `"..*"` and computing the distribution of gender reveals that girls' speech accounts for 46.934% of the words and boys' speech for 50.063%. The percentages within each row must therefore be weighted by taking into account the group size differences.

3 Quantitative analysis and visualization

The row percentages from the distribution in Corpuscle were downloaded and imported in Python. Values for unknown gender were dropped.

[1] These words were suggested by Erlend Astad Lorentzen and the idea for the exercise was suggested by Knut Hofland.

[2] http://clarino.uib.no/corpuscle

Word	Sum	f	m	-
	1041	518 49.76% 50.226%	489 46.974% 45.71%	34 3.266% 4.064%
beat	89	27 30.337%	62 69.663%	
bloke	87	29 33.333%	55 63.218%	3 3.448%
cat	37	23 62.162%	13 35.135%	1 2.703%
clothes	68	44 64.706%	21 30.882%	3 4.412%
cool	80	22 27.5%	54 67.5%	4 5%
crap	141	47 33.333%	91 64.539%	3 2.128%
football	48	11 22.917%	37 77.083%	
kiss	38	29 76.316%	7 18.421%	2 5.263%
love	191	129 67.539%	53 27.749%	9 4.712%
model	19	14 73.684%	2 10.526%	3 15.789%
music	89	38 42.697%	48 53.933%	3 3.371%
phone	154	105 68.182%	46 29.87%	3 1.948%

Figure 1: Distribution of gender relative to word (screenshot from Corpuscle).

The percentages for male and female were first weighted to reflect an ideal 50% division by multiplying boys' values by 50/50.063 and girls' values by 50/46.934. These numbers do not necessarily add up to 100% within each row, since the values for unknown gender were discarded. Therefore, the weighted percentages were further scaled so as to add up to 100%.

These final normalized percentages are presented in Table 1, which is sorted to show gender polarity better. It can be seen that these numbers are indeed slightly different from the unweighted percentages in the Corpuscle distribution in Figure 1. The same results are also depicted in the sorted stacked barplot in Figure 2.

Table 1: Percentages weighted by group sizes.

Word	f	m
football	24.1	75.9
cool	30.3	69.7
beat	31.7	68.3
crap	35.5	64.5
bloke	36.0	64.0
music	45.8	54.2
cat	65.4	34.6
clothes	69.1	30.9
phone	70.9	29.1
love	72.2	27.8
kiss	81.5	18.5
model	88.2	11.8

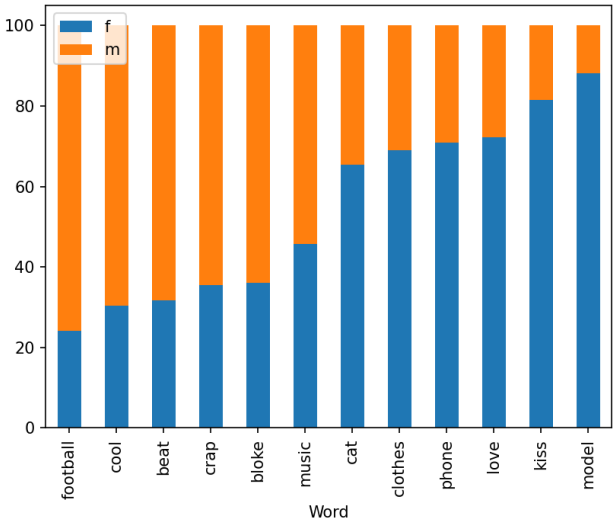

Figure 2: Percentages weighted by group sizes.

4 Discussion

This article has tested whether some words assumed to be characteristic for boys would occur at higher frequencies in boys' speech in COLT, and vice versa for girls. The present analysis has taken into account that there are group size differences in the corpus.

The extent to which the initial assumptions seem to be confirmed can be seen from the results, as shown in Table 1 and Figure 2. These results suggest that there are words which, to varying degrees, are more characteristic for one group than for the other. Any conclusions should however be considered with care because the statistical significance of each difference has not been calculated. The vocabulary sample in this study was very small and based on intuition; furthermore, the validity of the corpus data is dependent on the quality and representativeness of COLT.

5 Epilog

This paper is mainly intended as a pedagogical tool to illustrate how corpus data can be analyzed and visualized by a simple Python program which produces text, tables and plots which are subsequently input to LaTeX (Kopka and Daly, 2004; Mittelbach et al., 2004). It also demonstrates reproducibility. This research as well as the paper itself can be reproduced through the following steps:

1. Run the first part of the Python notebook, which defines the two lists of words and constructs the query. Optionally, choose different words.

2. Select the ICAME collection and the COLT corpus in Corpuscle; paste the query and search. Calculate the distribution of *gender* relative to *word*, type *absolute*. Download the fractions and put the downloaded file *distribution.txt* in the folder *data_path* as specified in the notebook. Also, save a new screenshot of the distribution table to *distribution.png* in the same folder as the present file.

3. Run the remainder of the Python notebook, which computes the weighted percentages and writes text, tables and plots to various files which are input to LaTeX.

4. Typeset the paper and proofread.

An advantage of such a workflow is that it can be replicated easily. Another advantage is that changes to the data, for instance, by selecting different words to be looked up, require minimal and fast updating.

References

Boulis, Constantinos and Mari Ostendorf (2005). "A Quantitative Analysis of Lexical Differences between Genders in Telephone Conversations." In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics – ACL '05*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 435–442.

Kopka, Helmut and Patrick W. Daly (2004). *Guide to LaTeX*. 4th ed. Boston: Addison-Wesley.

Meurer, Paul (2012). "Corpuscle – a New Corpus Management Platform for Annotated Corpora." In: *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by Gisle Andersen. Studies in Corpus Linguistics 49. Amsterdam/Philadelphia: John Benjamins, pp. 31–49.

Mittelbach, Frank, Michel Goossens, Johannes Braams, David Carlisle, and Chris Rowley (2004). *The LaTeX Companion*. 2nd ed. Addison Wesley.

Språkkontakt og ungdomsspråk i Norden (1999). *COLT – The Bergen Corpus of London Teenage Language (with Audio Recordings). Distributed by the CLARINO Bergen Centre*. URL: http://hdl.handle.net/11495/D9B6-13F8-41BB-1 (visited on 06/18/2021).

Stenström, Anna-Brita, Gisle Andersen, and Ingrid Kristine Hasund (2002). *Trends in Teenage Talk*. Studies in Corpus Linguistics 8. John Benjamins Publishing Company.