Friedrich Alexander Universität Erlangen-Nürnberg
Department of English and American Studies
Summer semester 2021
Pragmatics Seminar

Importance of vocabulary for categorizing academic texts

Hausarbeit

David Schwarzbeck

Informatik, Bachelor of Science

Contents

Abstract

This paper presents a computational approach of categorizing texts as academical as an alternative to rather complex approaches of text genre categorization with sophisticated mathematical and statistical tools like neural networks by only using simple linear discriminant analysis on the corpus data of the individual lexical items.
Since only the vocabulary of a text is investigated the conducted experiments also show the importance of a specific vocabulary for academia and the categorization of academic texts.
Additionally, the importance of the used word classes is investigated, which shows a preference towards certain classes in academia, but isn't statistically relevant enough to provide a categorization criteria.
The last experiment shows that investigating only the words from certain word classes i.e. using a word class filter yields better results than investigating all word classes, but the filter depends on if written, spoken or both text types are investigated.

In conclusion this paper shows, that vocabulary is an important part of academic papers, and they are categorizable by only looking at the used lexical items. For the purpose of this paper a program was written, which can be used to replicate the experiments or to rate the vocabulary of texts on its relation to academia, which can be quite useful to improve ones academic vocabulary (see section 3).

1 Introduction

Many people would argue that academical or scientific papers differentiate from novels, magazines, blog entries on the web etc. mainly by the method with which they were written: scientific papers need to rely on sources and related work to support the thesis and to connect to other recent research. Statements need to be proved by experiments or other sources and the intention of academic texts is mainly to propagate knowledge and new findings. This fact can be seen by looking into explanations or instructions for students on how to write a scientific paper (e.g. the paper *How to Write Your First Scientific Paper* Reis and Reis, 2013) which seldomly deal with aspects of language, but more with finding a topic to write about or with finding and working correctly with sources.

This may be because instructors, professors or tutors expect students to be already able to write in a scientific language. Nevertheless many instructions take the time to explain the basic structure, the needed constituents and even the right formulation for a title Reis and Reis (2013, p. 4-5), but not which language or which vocabulary to use. Does that mean that vocabulary is not important for academic texts?

The question which of the two following texts is academic and which is not can be quite easily answered:

> **Text 1:** The building loomed ahead of me, dark blue, almost black in the foggy night. Only the pale light from a single street light illuminating it. The streets were empty of all but two cars; to my left was a white van, missing a wheel and left for the scavengers of the street to pick clean. The side panels had been graffitied with the tags of the local artists. u/danwritesbooks (2021)

> **Text 2:** In a way, cohesion and coherence constitute different perspectives from which a text can be analyzed. The analysis of cohesion focuses on the text- internal links between linguistic units such as words and structures, whereas the investigation of coherence concerns such aspects whether a sequence of words and linguistic structures presenting certain facts, events or relations between them make sense. Herbst (2010, p. 285)

It is quite obvious that the second text is the academic one, yet neither of both excerpts contains any overall structure (there are no titles or sections) and no citations, which only leaves the claim that readers are capable of differentiating between academic and unacademic texts by their language i.e. the used vocabulary, collocations and other grammatical features.

For language learners, writing scientific or academic texts is a challenge - which makes knowing on which aspects to concentrate which vocabulary to use and how to formulate formal sentences important i.e. knowing which language aspects distinguish academic texts from other texts. Especially choosing the correct vocabulary can become quite difficult, since language learners usually lack the experience of knowing how words (or more specifically: lemmas) correlate to certain genres, especially to academia.

In the corpora studies and experiments, academia and academic texts will be treated as a from of text genre, corresponding to the categorization of words into different genres - including academia - by the *COCA* corpus.

This paper will further investigate the importance of vocabulary for the recognition of academic texts on basis of corpus data. For this purpose a program was written to determine an overall academic rating of a text based on the contained words and their corpus evidence in other academic texts.

2 Related Work and differentiation

Categorization of academic texts is a special case of text genre recognition. There are many important applications of text genre recognition like the sorting of texts by their genre (with special importance for academic texts e.g. for students who search for references or sources), for search engines or for linguists to investigate the abstract concept of ,,genres“, see Kessler et al. (1997, Introduction), hence there are many papers which dealt with that matter.
Most papers deal with a multitude of genres and with relative complex methods. The following generic cues were proposed by Kessler et al. (1997, chapter 2-3):

- Structural cues: e.g. part-of-speech tags
- Lexical cues: e.g. Latinate affixes are often used in scientific articles than in other texts
- Character-Level cues: ,,punctuation cues [...] used to mark text categories like phrases, clauses, and sentences“
- Derivative Cues: ,,Derivative cues are ratios and variation measures derived from measures of lexical and character-level features.“

The methodology includes Logistic Regression and even the use of neural networks in a second set of experiments on basis of corpus data from the brown corpus.

Nevertheless, this paper concentrates more on the categorization of texts into a single genre: academic or non-academic. For this categorization the feasibility of limited investigation with only the vocabulary (i.e. lexical cues) is examined and tested. Since this approach constitutes a far more limited and simpler method than proposed by Kessler et al. (1997) and just focuses on vocabulary without collocations or derivative or character-level cues, much of the proposed sophisticated methods will exceed the necessities for this question.

Therefore instead of logistic regression, linear discriminant analysis will be used, similar to the approach of Karlgren and Cutting (1994), which proposes a different set of parameters, including again derivative cues or character level ones, but also more lexical ones, which rather concentrate on occurrences of single words (e.g. ,,therefore“ count, ,,Me“ count) or the quantity of word classes (e.g. ,,Adverb“ count), see Karlgren and Cutting (1994, p. 1072 Table 2), which will be investigated in 5.1.

3 Resources

The used Corpus data for the statistical research and as included in the github project is a sample of the 5050 most frequent lemmas and 11461 corresponding word forms from the *COCA* Corpus freely available on the website www.wordfrequency.info. There are also samples of every tenth word of the top 60.000 lemmas, but since we are more interested in the most used words, the top 5050 lemmas are a more adequate sample.

The program which was developed for the purpose of this paper can be found at https://github.com/Frobeniusnorm/AcademicalPaperEstimator with a complete description and instructions on how to use it. It is completely open source and free to use (see License). It can be used to replicate the experiments, to further investigate the statistical methods used in this paper or to improve ones academic vocabulary for writing texts or papers by showing the least and most academic words and an overall rating of the text.

4 Methodology

4.1 Corpus Data

To investigate the correlation between vocabulary and the academic genre the categorization of lemmas into genres from the lemma list of the *COCA* corpus will be used. The lemma list contains five main genres: spoken, fiction, popular magazines, newspapers and academic and a large collection of sub-genres (see https://www.wordfrequency.info/uses.asp), which are not relevant for this research - only the academic genre will be taken into account.

The overall frequency and the frequency of each genre is given in two formats: the raw, absolute frequency and a normalized frequency given in occurrences per million words (overall or relative to the genre). Because the numbers of overall words in each genres differs, the normalized form will be used. Comparison across genres will not suffice for a good categorization, but its more sensible to compare the normalized frequency of a genre - in this case the academic genre - to the overall normalized frequency, which is also provided by the *COCA* corpus sample.

To accomplish this the program constructs a word database from the corpus data which consists of the lemma and a corresponding distribution value in academic texts and its word class (the *COCA* corpus assigns a word class to each word in the ,,part of speech" column, so the problem of words being part of multiple word classes was already dealt with by the corpus design). Only words from the word classes nouns, verbs, adjectives, adverbs, prepositions, pronouns, conjunctions and interjections are adopted, because the word classes will be further investigated later on and thus have to be limited to a selection. The distribution value is the ratio of the academic normalized frequency ($acadPM(w)$) and the overall normalized frequency ($perMil(w)$), which can be mathematically described for a word w by

$$d(w) = \frac{acadPM(w)}{perMil(w)}, d \in [0, \infty) \subset \mathbb{R} \tag{1}$$

So naturally 0 would mean the word does *not at all* occur in academic texts, 1 means that it occurs *as much* in academic texts as in other text genres and a value x with $x > 1$ would mean it occurs x times *more often* in academic texts, than it does in non-academic texts.

Additionally a second table is created that contains a mapping of a word form to its lemma as referenced in the word form table of the *COCA* corpus. Only the word class of the lemma is saved, not the word class of the word form (usually they don't differ, but if they do, the one of the word form is ignored).

4.2 Data processing and evaluation

To guarantee reasonable look up times, each word form and each lemma is saved by a hash value, which guarantees c.a. the same runtime independent to the size of the word lists. Since the distribution value d (further called *distribution score*) is a linear value, it can be averaged for a set of words. so the distribution value of a text t can be described by the function:

$$d(t) = \sum_{i=1}^{|W(t)|} \frac{1}{|W(t)|} \cdot d(W(t)_i) \tag{2}$$

Where $W(t)$ gives the set of words contained in the text t. Since this approach treats a text as a set of independent words, which is a strong simplification of the approach described by Kessler et al. (1997, chapter 2 - 3) , which additionally proposes e.g. ratios like words per sentence and therefor also take the relations of words to each other into account, it is important to find

a method to actually split a text into words (i.e. define the concept of words logically). This is done by defining a regex (*regular expression*) which can be used to split a text into a set of lexical items. Words like ,,we're" are split up into two words (,,we" and ,,'re"), because the *COCA* Corpus treats ,,'re" and similar abbreviations as individual items. The used regex splits a text into words at punctuation symbols (, . ; : & / -) and at whitespaces. The abbreviation ,,i.e." is discarded, because it would lead to the analyzation of the words ,,i" and ,,e", but ,,i" is already a word in the english language and would thus distort the result. The regex can be found in the file https://github.com/Frobeniusnorm/AcademicTextEstimator/blob/main/src/main/scala/TextEstimator.scala in the method *estimateAcademical*. All word forms and lemmas are transformed into lower case, as well as the words of each text.

4.3 Test data

To test the developed approach some test texts have been selected. The texts are pre-categorized as academic or non-academic and from as many genres as possible. They are from now on referenced in the data only by their abbreviations.

source	*abbreviation*	*selection*	*#words*	*academical*
Herbst (2010)	[acad1]	chapter 17.1	1075	yes
Brown et al. (2010)	[acad2]	Introduction	1833	yes
Rubio-Fernandez (2019)	[acad3]	chapter 4 and 5	1662	yes
Scott and Steele (2011)	[acad4]	p. 50-54 & p. 65-66	3334	yes
u/danwritesbooks (2021)	[reddit]	everything	3421	no
Sharma (2021)	[news]	everything	2041	no
Brandi (n.d.)	[cook]	everything	1147	no
Herbert (1965)	[fict]	excerpt chapter 1	3345	no

Consisting of a linguistic book for university students, a paper about quantum computers, a pragmatics and a politic history paper on the one hand and a short-story on reddit, a BBC news interview about the situation in afghanistan in 2021, a recipe from a cooking blog and an excerpt from a science fiction novel on the other the selection is filled with a broad field of genres and text types.

5 Experiments

5.1 Correlation of word classes to academia

This experiment tests the correlation of academic texts to the distribution of certain word classes. The investigation of word class counts - along with other parameters - for text genre categorization was proposed by Karlgren and Cutting (1994, p. 1072), but it failed to discuss the overall success of those parameters and which word classes correlated most to which genre. Hence, this experiment will further investigate the correlation of several word classes to academia and if it is significant enough to provide a categorization criteria.

For this no additional data to the previous mentioned already constructed database from the corpus is necessary. To measure the distribution of words from a single word class in academic texts we can again use the calculated *distribution score* of each word and take the average of all words from one word class which can be computed by the same formula which computes the *distribution score* of a text (see equation 2), but with the set of words from the word class instead of the set from a text. This is especially interesting because it makes sense that word classes like Interjections with words like ,,yeah" are arguably far less used in academical texts as they are in spoken language.

The results are given in the figure on the right side. The values are - as described in Methodology 4 - given as the ratio of the distribution in academic texts and the distribution in overall text of each word class. Meaning a preposition occurs in average 1.39 more often in academic texts than in general, while pronouns occur 0.47 more often i.e. $\frac{1}{0.47} \approx 2.128$ times less often. As predicted interjections don't occur in the investigated academic texts and pronouns hardly occur, which makes sense, because in general authors of papers do not speak about themselves but rather use the passive form (occurrences of types of pronouns also form 4 out of 20 parameters in Karlgren and Cutting (1994, p. 1027)).

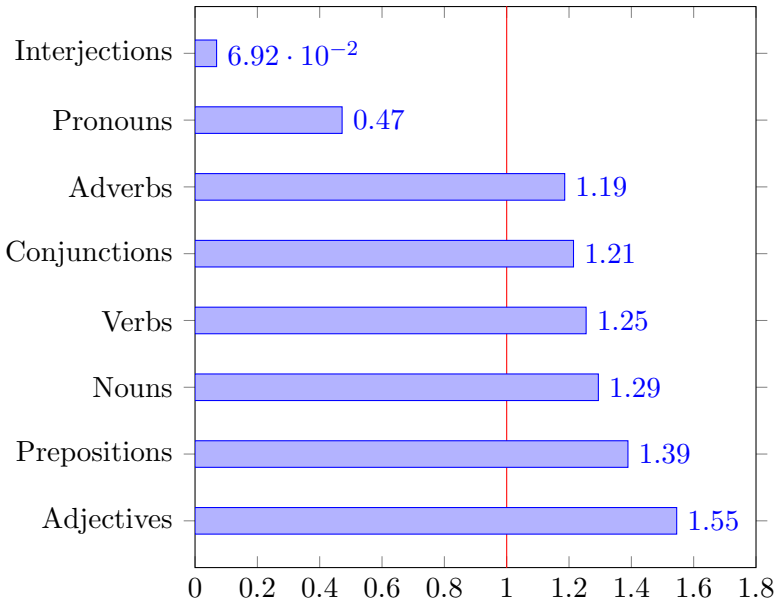


Figure 1: *distribution score* of each word class

This data already gives a good clue for writing scientific papers or texts: one should try to avoid interjections completely and pronouns as good as possible and rather concentrate on the other word forms, especially adjectives. Additionally, the fact that the distributions are rather extreme (e.g. pronouns and adjectives with 0.47 and 1.55) is quite interesting and raises the question if the information about the types and distributions of word classes in a text is already enough to categorize the text as either academic or unacademic.

To do this the program was further expanded to reduce a text to the word class frequencies of the contained words (see https://github.com/Frobeniusnorm/AcademicTextEstimator/blob/main/src/main/scala/TextEstimator.scala in the graphical user interface accessible through the button ,,compute only word classes“) and to take the average *distribution scores* corresponding to the *distribution scores* of each word class from figure 1.

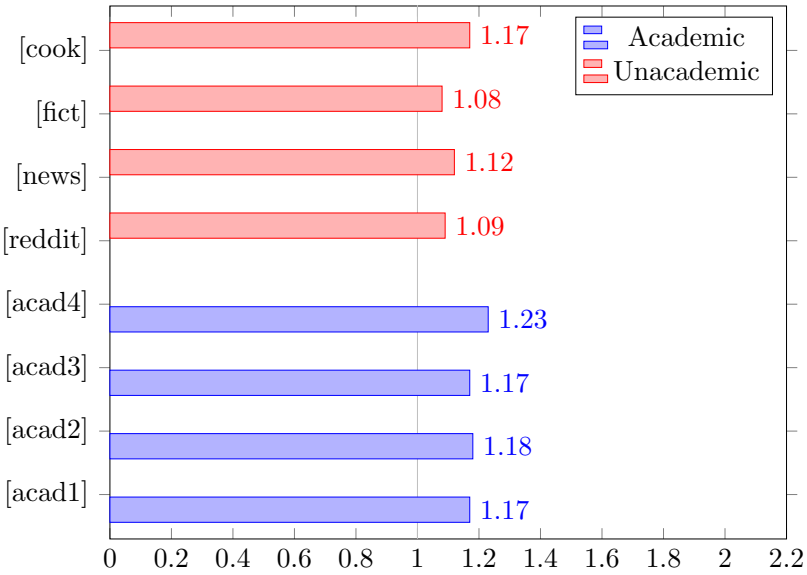


Figure 2: rating of texts by word classes

Although a small difference between academical and unacademic texts is visible (no academical text has a lower rating than an unacademic one) - so the difference seems to be indeed measurable - the relatively small difference of $1.23 - 1.08 = 0.15$ between the highest and lowest rating does not seem large enough to make word class distributions in a text a sufficient criterion. Additionally, no text has a real unacademic rating (no text has a rating below 1.0), which makes sense since the word class rating average is above 1.0 and words with a word class rating above 1.0 occur more often than the ones below, even in unacademic texts (e.g. the excerpt

from [fict] contains only 5 found interjections and 495 found pronouns, but 1170 found nouns, verbs, adjectives and prepositions).

Another reason may also be the actually used words. While pronouns in general have a lower *score* in average, the pronouns from the excerpt of [acad1] have a accumulated *distribution score* of 1.0 (in this case, of course, not the *distribution score* of the word class were accumulated but of each individual word), which shows that only counting the word classes leads to a loss of information about the choice of words.

In conclusion this experiment shows that for writing an academic paper it is recommendable to discard interjections and as many pronouns as possible, but the word classes in a text alone are not a big enough indication to clearly categorize a text as academic. This is in concordance to the paper *Recognizing Text Genres With Simple Metrics Using Discriminant Analysis*, which says: ,,the most important factors of the ones we testes are word length, sentence length and different derivatives of these two parameters“ Karlgren and Cutting (1994, p. 1073), so the investigated word classes did not bare much importance either.

5.2 Correlation of lexical items to academia

This experiment calculates the *distribution score* of the text by the method described in Methodology 4.2, i.e. by looking up the individual *distribution score* of each word and taking the average of all words. The results now show a clear differentiation of unacademic and academic texts: each text is recognized correctly (unacademic texts have a score below one and academic significantly above) and the differences are big enough to use the approach as a categorization criterion. A hard discrimination at the value of 1.0 between academic and unacademic texts, although it does not seem reasonable, because at a value of 1.01 or 0.99 it is not safe to clearly categorize the text in one of the binary states.

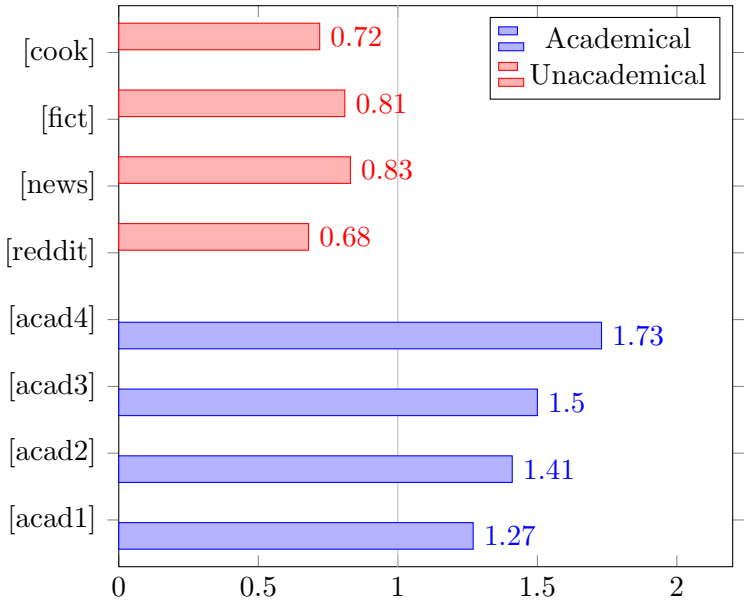


Figure 3: rating of texts by lexical items

rating	*from*	*to*
very unacademic	0.0	0.7
unacademic	0.7	0.875
neutral	0.875	1.125
academic	1.125	1.3
very academic	1.3	∞

Instead the value will be used to differentiate between 5 different ratings (see: left table). This should lead to a far better categorization since problematic border cases can be categorized as "neutral" and there are additional gradients for each category. In conclusion the estimation of the individual texts would be the following:

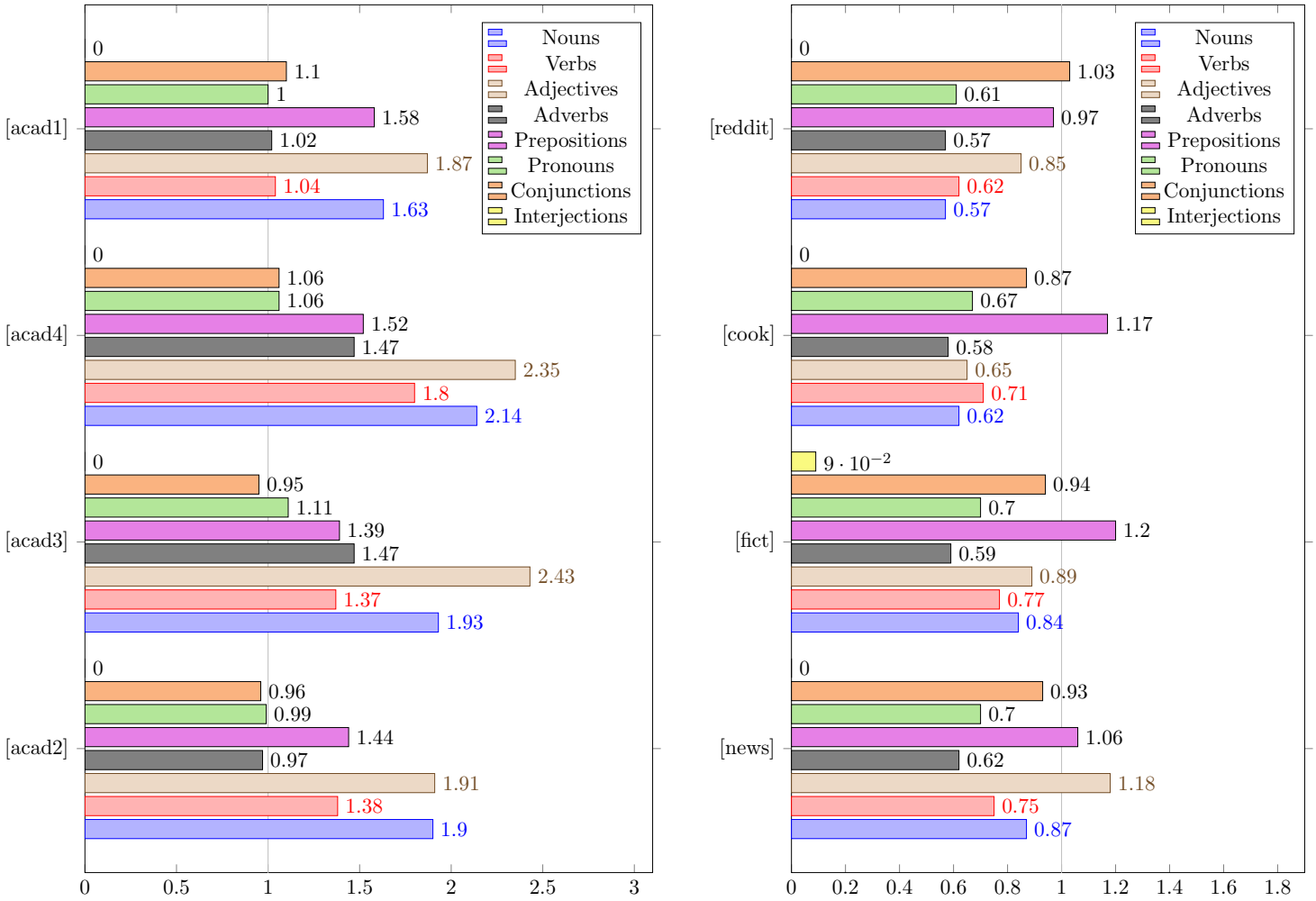
very unacademic: [reddit] *unacademic:* [news], [fict], [cook] *neutral:* -
academic: [acad1] *very academic:* [acad2], [acad3], [acad4]

This experiment shows quite clearly the strong connection between academia and the used vocabulary, therefor stresses the importance of a rich academic vocabulary for writing papers and proves that it alone is a sufficient criterion for categorizing texts as academic.

5.3 Investigation of word class filters

5.3.1 Ratings with a single word class

To investigate the correlation between experiment 5.1 and 5.2 a third experiment was conducted which again calculated the average *distribution score* of each text, but this time with only words from a single word class:



Interjections barely occur (only in one text), so they can mostly be ignored. Conjunctions show nearly no relation to the type of text (with a maximum variation of $1.1 - 0.87 = 0.23$ they don't show much difference at all) and are therefore also rather unimportant. For the other word classes one can measure the concordance to the text type by a value, which is simply calculated - mainly for the comparability to the other word classes - by the average of the average percentage of values over 1.0 in academic texts and the average percentage of values under 1.0 in non-academical text i.e.

$$con(wc, ac, uac) = \frac{1}{2} \cdot \left(\sum_{i=1}^{|ac|} \frac{1.0}{|ac|} \cdot (d(ac_i \cap wc) - 1.0) + \sum_{i=1}^{|uac|} \frac{1.0}{|uac|} \cdot (1.0 - d(uac_i \cap wc)) \right) \quad (3)$$

with wc = word class, ac = set of academic texts, uac = set of unacademic texts, $d(set \cap wc)$ = *distribution score* of the intersection of the words from a text with a certain word class (so only the selection of words, which are a member of that word class). This formula yields the following graph (see Figure 4), which shows little concordance of pronouns and prepositions and a strong concordance of nouns and adjectives. This raises the question if the results of Figure 3 are reproducible with only measuring nouns, adjectives and verbs.

5.3.2 Ratings with a word class selection

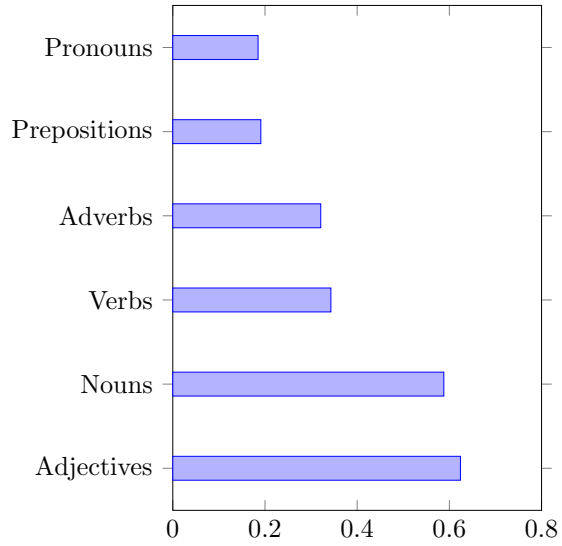


Figure 4: word class concordance by (3)

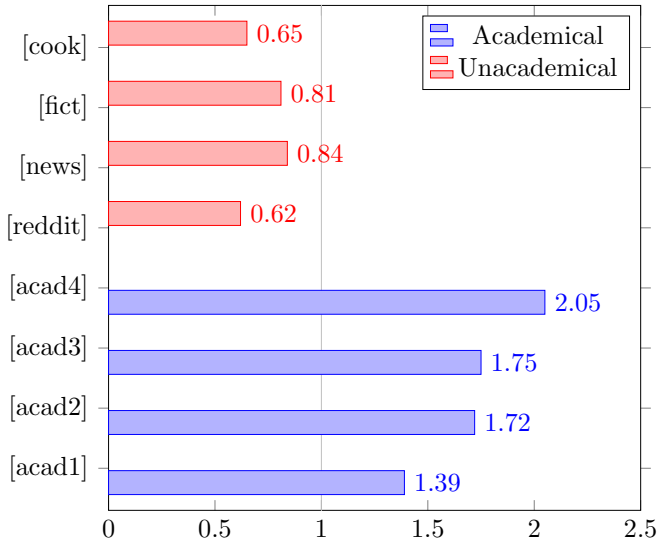


Figure 5: text ratings by nouns, verbs, adjectives

This approach actually shows better results than the one from 3: Except for the excerpt from [news], all texts are more clearly categorized i.e. the academic texts are rated higher, the unacademic lower. Additionally including Adverbs leads to overall marginally lower values, which benefits the categorizability in the unacademic texts, but contradicts it in the academic texts (see 6). Only using Adjectives and Nouns leads to overall higher values, which again contradicts the categorizability of unacademic texts to such a degree, that e.g. [news] can't be clearly rated unacademic any more (see 7). So the best results when rating or categorizing a text on how academic it is based on vocabulary are achieved by averaging the frequencies (or more specific for this case: the *distribution scores*) of all nouns, verbs and adjectives.

5.3.3 Additonal tests

source	*selection*	*number of words*	*estimation*	*result*
WHO (2009)	everything	5129	academic	1.79
Wikipedia-contributors (n.d.)	everything	2209	academic	1.22
Christensen and Jarosz (2014)	everything	6680	academic	1.53
Rais et al. (2011)	everything	1148	academic	1.52
McCusker (1992)	everything	7097	academic	1.61
Ellman (2002)	everything	4711	academic	1.5
Gross and Acquisti (2005)	everything	3424	academic	1.7
Shachaf and Hara (2010)	everything	2903	academic	1.75
Golding (1954)	The Sound of the Shell	2028	unacademic	0.68
Shakespeare (1597)	everything	9190	unacademic	0.77
BBC-News (2021)	everything	438	unacademic	1.16
Tompor (2021)	everything	648	unacademic	1.25
Rose (2021)	everything	1359	unacademic	0.85
Goddard (2021)	everything	212	unacademic	0.77
Common English Bible (2011)	Matthew 1 - 4	680	unacademic	0.79
Cahill (1997)	everything	3692	unacademic	0.81
Tolkien (1954)	everything	69887	unacademic	0.73
Harrigan (2018)	everything	2994	unacademic	0.99
Nasa-Science (2020)	everything	154	unacademic	0.85

Table 1: Additional texts rated by nouns, verbs and adjectives

The results from table 1 (academic ratings of the previously described approach applied to a larger set of texts) show a problem which did not arise in the previous test set, at least not clearly: professional news text have a rather academic vocabulary regarding nouns, verbs and adjectives. These values can be decreased by including pronouns (BBC-News (2021) $\mapsto$ 1.10, Tompor (2021) $\mapsto$ 1.17), but this again distorts the results of the other texts (see figure 4), BBC-News (2021) cannot be clearly categorized and Tompor (2021) is still rated academic. Nevertheless, it is partly reasonable, since a qualitative newspaper (in this case about politics and finance) is arguably of a more academic nature than other texts regarding language and vocabulary.

The unacademic text Harrigan (2018) was rated as neutral, although the writing style is clearly unacademic e.g. „Tick. Tock. Tick. Tock. Brrrrring! Analog clocks were, [...]“. This may be because of the subject specific topic (building clocks), which is mechanical and technical and therefore leads to the usage of specialized words, which are rated highly academical (e.g. „frequencies“ is rated 4.3, „scale“ 4.08, „components“ 3.88).

The unacademic text from Nasa-Science (2020) speaks against this explanation. The text explains galaxies in an educational manner for a younger audience, but was written by employees from NASA and therefore is still based on scientific explanations and uses subject specific words. According to the findings from rating Harrigan (2018) regarding specialized vocabulary, one would expect this text to be erroneously rated as academical, but it is clearly categorized as unacademic with a rating of 0.85. This may be because the corpus rates astronomic words less academic than mechanical and technical, maybe because of the occurrences in science fiction texts: the highest ratings were „shape“ with 2.18, „collection“ 2.15 and „system“ 2.12 which are about half as high as the highest ratings from Harrigan (2018). The provisional error rate is about $\frac{3}{27} \approx 11.1\%$ (this is not a fully representative error rate for this approach since the test set is still too small).

To further test and compare the approach with and without word class filters spoken and written texts from the American National Corpus (https://www.anc.org/data/oanc/contents) were tested. The last entry contains a set of fiction, prose and poetry texts from http://www.textfiles.com/stories/.

text set	type	*#texts*	*estimation*	*error rate 1*	*error rate 2*	*error rate 3*
plos	academic	252	academic	1.6%	0.0%	0.0%
charlotte	spoken	93	unacademic	1.1%	12.1%	3.2%
switch-board	spoken	2307	unacademic	2.3%	13.4%	5.3%
oup	non-fiction	45	academic	13.3%	2.2%	4.4%
textfiles	stories	287	unacademic	7.9%	4.8%	4.2%

Table 2: error rate 1: without filters, error rate 2: with filters, error rate 3: with filters

Error Rate 1 shows the results without word class filters (as in 5.2), error rate 2 shows the results with a word class filter consisting of nouns, verbs and adjectives (as in 5.3.2), the filter used for error rate 3 also includes interjections. The results show that spoken texts were better categorized without word class filters, in contrast to written texts, for which the word class filters showed better results. An explanation is that spoken texts contain more interjections and conjunctions instead of semantic bearing words since the required spontaneity leads to less information denseness, but the information about those word classes are lost with the first word class filter. The second filter is a compromise between spoken and written texts with including interjections and thereby achieves the best overall results.

In conclusion: when rating only written texts the best results are achieved when only including the nouns, verbs and adjectives, while when also investigating spoken texts interjections should be included too.

5.4 Problem cases

Although the experiments have shown promising results, there are a few simplifications and problems which may have distorted the results and therefore need to be mentioned:

- The sample of the *COCA* Corpus is rather small compared to the original list (5050 instead of 60000 lemmas), because of that many words are not recognized (keep in mind, that the statistic of the program does not actually only count items we consider as words: because of the regex mentioned in methodology 4.2 a string like "and - for this purpose" would be split into 6 words, although only 4 would be found). In the tested texts c.a. $\frac{3}{4}$ of the words are recognized, which is still a good amount considering the small sample.

- The *COCA* word form and lemma lists can cause some problems by their structure. E.g. in the text *Sponsoring Democracy: The United States and Democracy Aid to the Developing World*, Scott and Steele (2011) the word „US" occurs as an abbreviation for „United States", but is being treated as the pronoun „us" with a low *distribution score* of 0.39 and thus distorts the result.

- Seeing a text as a set of words can lead to problems, because it does not account for citations or examples. E.g. the excerpt from *English Linguistics: A Coursebook for Students of English*, Herbst (2010, chapter 17.1) has an academic rating of 1.27 without examples, but it only scores 1.22 with examples, however those examples cannot really be seen as part of the texts vocabulary.

The problem of ignoring constructs like idioms, collocation or grammatical features is explicitly **not** a problem of this approach, since they have been intentionally omitted to show the expressiveness of vocabulary.

6 Conclusion and discussion

The approach proposed by this paper i.e. averaging the *distibution scores* of all nouns, verbs and adjectives shows promising results for the recognition of academic texts. The findings from the extended test sets showed that the approach in general is reliably able to recognize academic and unacademic texts, but in cases where the texts' vocabulary is specialized or if it is written on a serious and professional language level (e.g. news text) it tends to be rated falsely.

The error rate is computable by averaging the error rates from the initial and extended text set proposed in this paper (consisting of 27 texts) and the 3 written sets from the American National Corpus: $\frac{1}{4} \cdot \left(\frac{3}{27} + 0.0\% + 2.2\% + 4.8\%\right) \approx 4.5\%$ (excluding spoken data). The success of this approach although would probably not be reproducible for multiple genre categorization since the paper Karlgren and Cutting (1994, p. 1072) which uses a similar approach, but with more parameters has shown that with only 2 categories the error rate is similar low $\frac{22}{500} = 4.4\%$, but if applied to 4 categories much higher $\frac{134}{500} = 26.8\%$. Presumably the lexical approach described in this paper would scale just as poorly.

The paper *Automatic Detection of Text Genre*, Kessler et al. (1997) did not list an experiment with two categories, but included the category „scitech" which is the one that comes closest to academia: 17% of the scitech texts were wrongly categorized as non fiction. Only 6% of the non-fiction texts were categorized as scitech, no other category was wrongly categorized scitech, not even editorial or reportage texts, see Kessler et al. (1997, p. 37, Table 3), which come closest to the category „news" in the coca corpus, therefore the approach of that paper did not reproduce the error cases found in the experiments from 5.3.3.

The lexical approach from 5.3.2 achieves better categorization of academic texts than by Kessler et al. (1997), which is a rather unfair comparison since Kessler et al. (1997) included 6 genres, the purpose of the proposed approach was only to categorize texts into 2 genres.

Nevertheless, the algorithm is in $\mathcal{O}(n)$, with n being the number of words in the investigated text and therefore much faster than using a neural network or including collocations or other word combination parameters. It therefore describes a fast way of categorizing academic texts, in costs of a small amount of quality, which make its application more fitting for e.g. search engines, were a rough estimation of the academicness of a text or site would suffice, and the requirement rather lies on fast categorization times, or for rating or self estimation of student papers in terms of vocabulary: the program shows the least academic words and therefore allows an easy way for replacing certain words to improve the overall language of a paper. Additionally, the approach proves the fundamental importance of vocabulary for academia, because all academic texts were rated correctly, only unacademic ones achieved an erroneous rating, which demonstrates that all academic texts share a common type of vocabulary, only unacademic texts do not.

7 Appendix

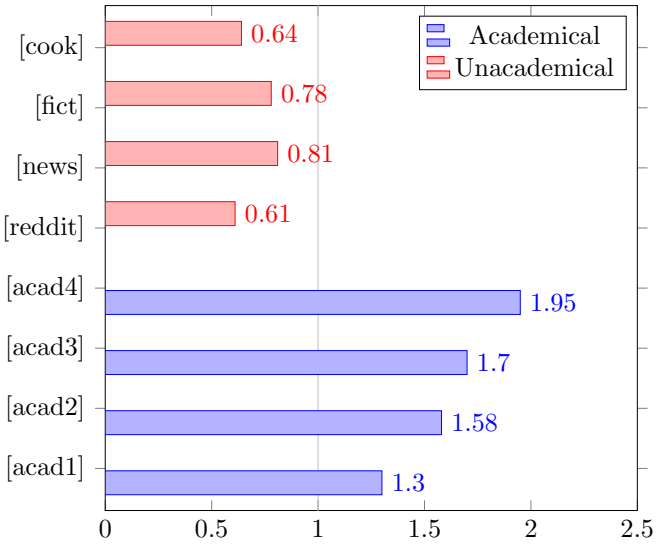


Figure 6: text ratings by nouns, verbs, adjectives and adverbs

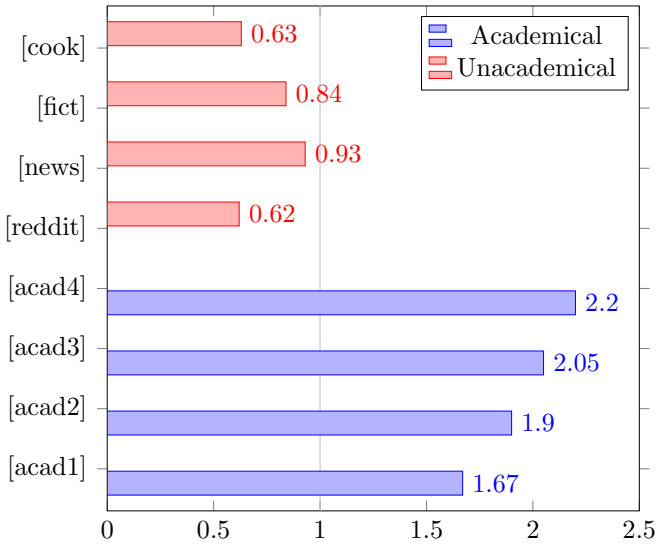


Figure 7: text ratings by nouns and adjectives

8 Bibliography

8.1 Resources

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, 1071–1075. https://doi.org/10.3115/991250.991324

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. *ArXiv, cmp-lg/9707002*.

Reis, S., & Reis, A. (2013). How to write your first scientific paper. *IEDEC*.

8.2 Excerpts for testing

BBC-News. (2021). Texas abortion: Biden vows 'whole-of-government' response to new law. *British Broadcasting Corporation*. https://www.bbc.com/news/world-us-canada-58424249, accessed on 2021-09-03.

Brandi. (n.d.). Easy vegan breakfast burritos (9 ingredients!) https://thevegan8.com/easy-vegan-mexican-breakfast-burritos/, accessed on 2021-08-25.

Brown, K. L., Munro, W. J., & Kendon, V. M. (2010). Using quantum computers for quantum simulation [Introduction]. *Entropy*.

Cahill, T. (1997). Forbidden. *Outside*. https://www.outsideonline.com/outdoor-adventure/environment/rocky-mountain-wildflowers-field-guide-nature-poetry/, accessed on 2021-09-03.

Christensen, P. H., & Jarosz, W. (2014). The path to path-traced movies. *Foundations and Trends in Computer Graphics and Vision Vol. 10, No. 2, p. 103–175*. https://graphics.pixar.com/library/PathTracedMovies/paper.pdf, accessed on 2021-09-02.

Common english bible. (2011). Christian Resources Development Corporation.

Ellman, M. (2002). Soviet repression statistics: Some comments. https://web.archive.org/web/20121122071204/http://artukraine.com/old/famineart/SovietCrimes.pdf, accessed on 2021-09-03.

Goddard, J. (2021). How are you feeling? *CUP of JO*. https://cupofjo.com/2021/03/how-are-you-feeling/, accessed on 2021-09-03.

Golding, W. (1954). *Lord of the flies*. Faber; Faber.

Gross, R., & Acquisti, A. (2005). Information revelation and privacy in online social networks (the facebook case). *Wayback Machine, pre-proceedings version*.

Harrigan, D. (2018). How to make a clock. *Ponoko*. https://www.ponoko.com/blog/how-to-make/make-a-clock/, accessed on 2021-09-02.

Herbert, F. (1965). *Dune* [chapter 1]. Chilton Books.

Herbst, T. (2010). *English linguistics: A coursebook for students of english* [chapter 17.1]. De Gruyter Mouton.

McCusker, J. J. (1992). How much is that in real money? a historical price index for use as a deflator of money values in the economy of the united states. *American Antiquarian Society*.

Nasa-Science. (2020). What is a galaxy? https://spaceplace.nasa.gov/galaxy/en/, accessed on 2021-09-04.

Rais, M., Khan, M. Z., Abbass, D., Akber, G., Nawaz, R., & Saeed-ul-Islam. (2011). A qualitative study on wildlife of chotiari reservoir, sanghar, sindh, pakistan. http://zsp.com.pk/237-247%20(4)%20PJZ-243-10.pdf, accessed on 2021-09-02.

Rose, C. (2021). Bears, berries, and unforgettable vistas: Hiking glacier national park. *USA TODAY TRAVEL*. https://eu.usatoday.com/story/travel/experience/america/national-parks/2021/08/23/glacier-national-park-montana-offers-majestic-views-and-brown-bears/8199459002/, accessed on 2021-09-03.

Rubio-Fernandez, P. (2019). Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity [chapter 4-5]. *Wiley Periodicals, Inc.*

Scott, J. M., & Steele, C. A. (2011). Sponsoring democracy: The united states and democracy aid to the developing world [p.50-54 & p.65-66]. *International Studies Quarterly*.

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science,*

Shakespeare, W. (1597). *Romeo and juliet*.

Sharma, N. (2021). A kabul evacuee's tale: Dread, chaos and the taliban. *BBC News, Delhi*. https://www.bbc.com/news/world-asia-58318374, accessed on 2021-08-25.

Tolkien, J. R. R. (1954). *The fellowship of the ring*. George Allen & Unwin.

Tompor, S. (2021). Social security retirement benefits could see big hike in 2022: What to know. *USA TODAY MONEY*. https://eu.usatoday.com/story/money/2021/09/02/social-security-cola-increase-2022-when-to-claim-benefits/5694096001/, accessed on 2021-09-03.

u/danwritesbooks. (2021). [hr] regret. *Reddit, r/shortstories*. https://www.reddit.com/r/shortstories/comments/p94673/hr_regret/, accessed on 2021-08-23.

WHO. (2009). Safety issues in the preperation of homeopathic medicines. *World Health Organization Press*. https://www.who.int/medicines/areas/traditional/Homeopathy.pdf, accessed on 2021-09-02.

Wikipedia-contributors. (n.d.). Apple. *Wikipedia The Free Encyclopedia*. https://en.wikipedia.org/wiki/Apple, accessed on 2021-09-02.