# Authorship Profiling Without Using Topical Information
# Notebook for PAN at CLEF 2018

Jussi Karlgren[1,2], Lewis Esposito[3], Chantal Gratton[3], and Pentti Kanerva[4]

[1] Department of Theoretical Computer Science, KTH, Stockholm
[2] Gavagai, Stockholm
[3] Department of Linguistics, Stanford
[4] Redwood Center for Theoretical Neuroscience, UC Berkeley

**Abstract** This paper describes an experiment made for the PAN 2018 shared task on author profiling. The task is to distinguish female from male authors of microblog posts published on Twitter using no extraneous information except what is in the posts; this experiment focusses on using non-topical information from the posts, rather than gender differences in referential content.

## 1 The PAN 2018 Authorship Profiling Experiment

This paper describes an experiment made for the PAN 2018 shared task on author profiling. The task is to distinguish female from male authors of microblog posts published on Twitter using no extraneous information except what is in the posts. The full task allows for using both images and text of the posts which are given in three languages: in this experiment we have only made use of the English-language material, and only the text. The training material consists of 1500 female and 1500 male authors, with 100 posts each. Microblog posts are short and these consist on average of $X$ words and $Y$ sentences [25,20].

## 2 What People Have Thought About Male And Female Language And Why

Robin Lakoff's 1973 book *Language in Woman's Place* [14] initiated conversations surrounding the role of gender in linguistic practice. While her work might better be described as a collection of ideologies of gendered language rather than an accurate depiction of men's and women's linguistic styles, it nonetheless cemented the legitimacy and significance of language and gender studies in its own right. And indeed, ideologies of how men and women differ in their use of language still pervade public discourse; and these discourses taint research from various disciplines that make bold claims about the gendered use of language without taking gender as a serious social construct worth investigating.

Perhaps one of the biggest myths about men's and women's is that women talk more than men, and the ubiquity of this belief has led researchers in fields tangential to linguistics to look for biological causes ([3, e.g]) despite the fact that such research is unsupported by quantitative data. Indeed, some work has found that there are no differences in the amount of speech men and women produce, such as Mehl's and colleagues' research that equipped male and female university students in the US and Mexico with microphones for several days, which randomly recorded them at various intervals [17]. Other work has found that men actually speak more than women, particularly in formal and task-oriented activities [9], and even young boys outstrip their female classmates, speaking three times as much and calling out answers 8 times more [22].

Another common ideology about language differences among women and men is that women use more hedges than men. This idea generally arises from the folk ideology that women tend to be less sure of themselves. But just as the quantitative evidence described above doesn't support the "talkative women" ideology, work on hedges has similarly found that men and women use hedges at comparable rates [19]. Similarly, the notion that women also use other linguistic features that signal low confidence, like creaky voice and innovative like, isn't supported by the data either. Men and women have been shown to use creaky voice at roughly equal rates [1], and the same holds for different discourse functions of like [5].

But none of this is to say that men and women don't participate in linguistic practices in unique ways. The ideologies described above are exactly that: ideologies, rooted in bias and lacking quantitative reality. Quantitative sociolinguists nonetheless consistently find broad gender patterns in the use of linguistic features. Women, more often than not, drive vocalic sound change [13], leading men in the use of incoming variants. Searching for biological or essentialist motivations is an untenable approach, as male-led sound changes have indeed been documented ([18, e.g.], ruling out the potential for sex-based effects on linguistic production. For this reason, Eckert urges us to consider the kinds of social milieu that men and women occupy in society [7]. As men have historically enjoyed greater power than women in all domains of public and private life, and given that they have been deprived of social and political capital, women may have greater motivation to make use of various kinds of symbolic capital. It should thus not be surprising that women, in the aggregate, are more advanced than men in innovative phonological changes that, at their inception, are believed by many sociolinguists to be imbued with socio-symbolic meanings [8]. Beyond components of sound change, there are no doubt other linguistic features that men and women employ variable, but the perhaps more interesting question for scholars is why these differences exist, and for whom do they not exist.

## 3 Features and Variables of Interest

In the present data set, where the gender of authors can be expected to be distinguishable with a precision of around 80% using largely lexical cues [21]. Lexical variation is highly determined by topic, and essentially much of the results can be reduced to the observation that female and male authors write about different things: many discourse topics are strongly gendered.

If the task is to distinguish female and male authors in this specific data set or very similar ones from more or less the same time period, a well trained topical detector will be useful. If the task is to detect what differences may be systematic between genders across topics and over time, topic will be less reliable as a gender maker. Our experiments start from the assumption that topic is a confounding and non-sustainable variable for the general case. We also wish to point out that for many downstream tasks, the distinction between male and female author may be less useful than other stable characteristics, and that as in many classification tasks, assuming that the number of classes is fixed a priori may lower both the reliability and the usefulness of the classification.

### 3.1 Linguistic Processing

We process the linguistic data in a vector space model which incorporates lexical linguistic items together with constructional linguistic items in a unified computational framework.

**Vector Space Models for Meaning** Vector space models are frequently used in information access, both for research experiments and as a building block for systems in practical use at least since the early 1970's [23,6]. Vector space models have attractive qualities: processing vector spaces can be done in a manageable implementational framework, they are mathematically well-defined and understood, and they are intuitively appealing, conforming to everyday metaphors such as "near in meaning" [24]. The vector space model for meaning is the basis for most all information retrieval experimentation and implementation, most machine learning experiments, and is now the standard approach in most categorisation schemes, topic models, deep learning models, and other similar approaches. In this experiment we encode each post of each author into a vector, and use those vectors to represent the authors profile.

**Construction Grammar** The *Construction grammar* framework is characterised by the central claims that linguistic information is encoded similarly or even identically with *lexical items*—the words—and their *configurations*—the syntax, both being linguistic items with equal salience and presence in the linguistic signal. The parsimonious character of construction grammar in its most radical formulations [4, e.g] is attractive as a framework for integrating a dynamic and learning view of language use with formal expression of language structure: it allows the representation of words together with constructions in a common framework. For our purposes construction grammar gives a theoretical foundation to a consolidated representation of both individual items in utterances and their configuration. In this experiment, after dependency analysis of each sentence of each post, features of potential interest in each sentence are extracted to represent the sentence together with some of its lexical items.

## 4 Technical Description

To represent authors by features of their posts as vectors, we use a high-dimensional model based on *random indexing* [10]. The idea is to compute with high-dimensional vectors [11] using operations that do not modify vector dimensionality during the course of operation and use. We use 2,000-dimensional vectors in these demonstrations and experiments. Information encoded into a vector is *distributed* over all vector elements. Computing begins by assigning *random seed vectors* or *index vectors* for basic objects. In working with text each observed word and each observed construction of interest in the collection can be represented by an index vector consisting of 0s, 1s and $-1$s. These can easily be generated on the fly if new lexical or constructional items appear during processing. Index vectors remain unchanged throughout computations. Typically, index vectors are sparse, and in our model have 10 non-zero elements with an equal number of 1s and $-1$s. Each item also is given a *context vectors*, where observations of cooccurring items are recorded through *vector addition*, and if necessary, *vector permutation* , which reorders (scrambles) vector coordinates. These operations are inexpensive computationally and allow for a very large feature space within a bounded memory footprint. As in most similar models, *vector similarity* is measured by cosine between the vectors, with values between $-1$ and and 1 [12].

## 5 Representation of Posts

The posts were segmented into sentences and word tokens using NLTK [2], and each token tagged by Penn Treebank lexical category using CoreNLP [16,15]. The sentences were further analysed for syntactic dependencies, again using CoreNLP.

### 5.1 Full Text Baseline

As a baseline, all words of each post is included in the representation. Each word was assigned a random index vector and added into the representation weighted by logarithmic frequency weighting to damp the relative effect of highly frequent words and increase the weight of infrequent ones. This weighting scheme was not optimised especially for this material.

A quick glance through the lexical date will show that some words are more often typically used by female than male authors. The numbers in Table 1 are taken directly from the vector space model. The proportion of female and male authors in the 100 authors closest to each word in the vector space is given, along with their frequency in the entire training collection.

Some terms (*game*, *win*, *birthday*) can fairly be called topical. Others reflect more stylistic or attitudinal usage (*happy*, *love*, *wrong*, *sure*). Terms such as *stuff*, while referential, simultaneously reveal volumes about the authors attitude to the topic under treatment. How to establish that cline of referentiality or topicality vs attitude is a research challenge which partially could be addressed using measures from search technology.

### 5.2 POS sequences

Each sentence was represented as a sequence of Penn Treebank POS labels. These labels are not always well chosen, but no correction of the output of the NLTK tagger was done. Subsequences of length three were extracted for each sentence.

(1) a. Anyone have a travel rest pillow I could borrow for a long trip?

| | frequency | ♂ | ♀ |
|---|---|---|---|
| sure | 2 537 | 69 | 31 |
| wrong | 1 369 | 33 | 67 |
| hope | 3 516 | 29 | 71 |
| life | 2 597 | 28 | 72 |
| game | 2 019 | 70 | 30 |
| team | 1 611 | 64 | 36 |
| win | 1 919 | 65 | 35 |
| America | 1 288 | 62 | 38 |
| birthday | 1 097 | 33 | 67 |
| happy | 1 952 | 37 | 63 |
| love | 5 216 | 33 | 67 |
| stuff | 1 078 | 63 | 37 |
| fun | 1 309 | 19 | 81 |
| thank | 4 183 | 27 | 73 |
| thanks | 3 185 | 29 | 71 |
| women | 1 332 | 32 | 68 |
| Yes | 1 423 | 38 | 62 |
| amazing | 1 859 | 21 | 79 |

**Table 1.** Examples of lexical skewness in the data

b. NN, VBP, DT , NN, NN, NN, PRP, MD, VB, IN, DT, JJ, NN, "."
c. [[NN, VBP, DT] , [VBP, DT, NN], ... ]

One random permutation $\Pi$ was generated for each POS label. One random vector $\boldsymbol{pos}$ was generated for encoding all POS labels. Each triple was represented by taking the POS vector and passing it through the POS permutations for the POS labels of the triple. All resulting triple vectors were then added into the post representation. This representation preserves the sequence of POS labels without conflating them for each position in a triple.

For example, the sequence DT, JJ, NN will be encoded as

$$S(DT, JJ, NN) = \Pi_{NN}(\Pi_{JJ}(\Pi_{DT}(\boldsymbol{pos}))) \qquad (1)$$

### 5.3 Constructional Elements

Some interesting observations can be made from a more general view of the terminological variation and some hypotheses about both syntactic and stylistic and attitudinal variation. Table 2 gives some statistics for some observable aggregate features of interest. Some of these are based on lists of lexical items of similar distributional and attitudinal qualities used in various sentiment analysis tasks; others are based on features extracted from dependency analyses from the Stanford CoreNLP package [15].

Amplifiers in general are slightly more prevalent in posts by female authors, but this separates interestingly with type of amplifier. Amplifiers can be separated into grade amplifiers (*very, extremely, ...*), veracity amplifiers (*truly, really, ...*), and anomaly amplifiers (*surprisingly, amazingly, ...*). The surprise amplifiers are what carry most of the difference between female and male authors.

First person singular personal pronouns (*I, me, myself, my, mine*) are used more by female authors than male authors. *We* and its inflected forms, by contrast, are evenly distributed.

Profanity is used more by male authors; interjections (*lol, omg, hey, oh, wtf, ...*) more by female authors.

Some verbal constructions are skewed: male authors use more passives; female authors more progressive tense. Modal auxiliaries are used more by male authors to a certain extent, and this coupled with the observation that male authors also use more hedges and downtoners can most likely be traced to differences in which discourses male and female authors engage in: male authors appear to more often be participate in political debates and argumentation compared to female authors.

| | |
|---|---|
| all amplifiers | 43 57 |
| grade amplifiers | 47 53 |
| anomaly amplifiers | 36 64 |
| veracity amplifiers | 42 58 |
| hedges and downtoners | 74 26 |
| uncertainty | 64 36 |
| p1 singular | 17 83 |
| p1 plural | 53 47 |
| p2 | 37 63 |
| p3 | 59 41 |
| profanity | 69 31 |
| interjection | 37 63 |
| passive constructions | 67 33 |
| progressive tense | 40 60 |
| should | 61 39 |
| would | 72 28 |
| could | 56 44 |
| think and cogitation verbs | 66 34 |
| utterance verbs | 67 33 |
| love terms | 15 85 |
| hate terms | 43 57 |
| boredom terms 59 | 41 |
| dislike terms | 56 44 |

**Table 2.** Examples of complex feature skewness in the data

These and other similar features (tense of main verb, definiteness of subject and object, various categories of adverbials of place, time, and manner) are each encoded with a random index vector and, in keeping with the constructional grammar principles mentioned above, included in the representation as if it were a lexical item.

### 5.4 Generalised Lexical Elements

To reduce the topical content nouns, verbs, and adjectives are replaced with their corresponding POS tag, using the Penn tagset. This means adjective comparation, verb tense, and noun number is preserved, but the actual referential meaning of the word will have been taken out.

(2) a. Anyone have a travel rest pillow I could borrow for a long trip?
b. NN, VBP, a , NN, NN, NN, I, could, VB, for, a,, JJ, NN, "."

### 5.5 Centroids and Pool Depth

As a final series of representational parameter choices, given a vector space of sentences along the lines above, we must first determine if a (1) post is best represented as an average, or a vector centroid, of its constituent sentence vectors or as a bag of separate vectors; if (2) an author is best represented as an average, or a vector centroid, of its constituent post or sentence vectors or as a bag of separate vectors; and (3) if a gender is best represented as an average, or a vector centroid, of its constituent sentence, post or author vectors or as a bag of separate vectors. We have here elected to use an author centroid for each author comprised of a sum of post vectors, in turn comprised of a sum of sentence vectors, but not to average the authors into a single gender vector.

Given such an author space and a new author of unknown gender with a vector in the space, the next question is to decide how to assess its position in author space. We can assign the author the same gender as its nearest neighbour in space or use a broader range to pool a number of neighbours. In the following tables, we show results from using only the very nearest neighbour and from the 11 closest neighbours.

Both these questions — centroids or bags of vectors, and how to assess position in author space, are amenable to further experimentation and attendant improvement using classification algorithms of various levels of sophistication.

## 6 Cross Validation Results on the Training Data

All training sentences, posts, and authors are encoded as vectors using all the above features. The nature of the representation is such that these overlayed encodings of multiple features can be used fully or with only some of the features in play. Test sentences, posts, and authors are encoded with all or some subset of the features, and the classification is done using simple cosine calculation to find the closest neighbour to the test author in question.

Tables 3 and 4 give a combined picture of the quality of the various features sets — all words (WDS), generalised content words (NON-TOPIC), part of speech triples (POS), constructional features (CXG), together and separately. The results given are based on 3-fold cross-validation over the training data. The submitted run is based on the -WDS condition, using all feature types except content words, and at a pool depth of 1. Notable from the results is that precision for the female authors is greater (at an attendant cost to recall). This gives us reason to believe that the representation of female authorship in this space is different than that of male authorship. One tentative but likely explanation is that there are more than two styles, and that there are more female styles than male styles among them in this material.

| gender | ♀ | ♂ | both | ♀ | ♂ | both |
|---|---|---|---|---|---|---|
| pool depth | | 1 | | | 11 | |
| WDS | 0.68 | 0.5 | 0.5867 | 0.66 | 0.46 | 0.5400 |
| NON-TOPIC | 0.5 | 0.49 | 0.4933 | 0.6 | 0.53 | 0.5533 |
| POS | 0.52 | 0.45 | 0.4867 | 0.53 | 0.46 | 0.4933 |
| CXG | 0.59 | 0.48 | 0.5267 | 0.53 | 0.44 | 0.4933 |
| ALL | 0.75 | 0.52 | 0.6067 | 0.76 | 0.49 | 0.5733 |
| -WDS | 0.54 | 0.46 | 0.4867 | 0.67 | 0.5 | 0.5333 |
| -NON-TOPIC | 0.52 | 0.47 | 0.4933 | 0.63 | 0.54 | 0.5800 |
| -CXG | 0.58 | 0.55 | 0.5600 | 0.69 | 0.58 | 0.6133 |
| -POS | 0.59 | 0.5 | 0.5267 | 0.71 | 0.55 | 0.6000 |

**Table 3.** Accuracy for cross-validation runs on the training data

| gender | ♀ | ♂ | ♀ | ♂ |
|---|---|---|---|---|
| pool depth | 1 | | 11 | |
| WDS | 0.56 | 0.44 | 0.63 | 0.68 |
| NON-TOPIC | 0.39 | 0.34 | 0.59 | 0.77 |
| CXG | 0.45 | 0.45 | 0.62 | 0.51 |
| POS | 0.52 | 0.49 | 0.44 | 0.5 |
| ALL | 0.49 | 0.40 | 0.77 | 0.82 |
| -WDS | 0.36 | 0.27 | 0.64 | 0.84 |
| -POS | 0.34 | 0.41 | 0.73 | 0.82 |
| -CXG | 0.49 | 0.43 | 0.63 | 0.80 |
| -NON-TOPIC | 0.49 | 0.49 | 0.49 | 0.67 |

**Table 4.** Recall for cross-validation runs on the training data

## 7 What Does It All Mean

These are initial explorations to establish stylistic and attitudinal differences between categories of author. We believe that it would be more functionally appropriate to work with a broader palette of categories than two sexually determined categories; that topical variation majorises gender variation; that gender variation largely is socially determined in ways that has been studied extensively in sociolinguistics; that the intrinsic differences between categories invites further study of the variational space; that the signal found in these data could be better accommodated as an encoding to a more competent classifier; and that constructional analysis can be a key to a computationally habitable combination of lexical and syntactic analysis pipeline. We also acknowledge that none of these issues have fully been explored in this present experiment.

**Acknowledgements** Jussi Karlgren's work was done as a visiting scholar at the Department of Linguistics at Stanford University, supported by a generous VINNMER Marie Curie grant from VINNOVA, the Swedish Governmental Agency for Innovation Systems.

## References

1. Becker, K., ud Dowla Khan, S., Zimman, L.: Creaky Voice in a diverse gender sample: Challenging ideologies about sex, gender and creak in American English. New Ways of Analyzing Variation 44 (2015)
2. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media (2009)
3. Brizendine, L.: The Female Brain. Morgan Road Books (2006), https://books.google.com/books?id=-tpoFcql0kgC
4. Croft, W.: Radical and typological arguments for radical construction grammar. In: Östman, J.O., Fried, M. (eds.) Construction Grammars: Cognitive grounding and theoretical extensions. John Benjamins, Amsterdam (2005)
5. D'Arcy, A.: Like and language ideology: Disentangling fact from fiction. American speech 82(4), 386–419 (2007)
6. Dubin, D.: The most influential paper Gerard Salton never wrote. Library Trends 52(4), 748–764 (2004)
7. Eckert, P.: The whole woman: sex and gender differences in variation. Language Variation and Change 1, 245–267 (1989)
8. Eckert, P.: Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. Annual review of Anthropology 41, 87–100 (2012)
9. James, D., Drakich, J.: Understanding gender differences in amount of talk: A critical review of research. (1993)
10. Kanerva, P., Kristoferson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of the Cognitive Science Society. vol. 1 (2000)
11. Kanerva, P.: Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. Cognitive Computation 1(2), 139–159 (2009)
12. Karlgren, J., Kanerva, P.: Hyperdimensional utterance spaces—a more transparent language representation. In: Proceedings of Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, (DESIRES) (2018)
13. Labov, W.: Principles of Linguistic Change, Social Factors. Principles of Linguistic Change, Wiley (2001), https://books.google.com/books?id=LS_Ux3CEI5QC
14. Lakoff, R.: Language and woman's place. Language in Society 2(1), 45–80 (1973), http://www.jstor.org/stable/4166707
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), http://www.aclweb.org/anthology/P/P14/P14-5010

16. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313–330 (1993)
17. Mehl, M.R., Vazire, S., Ramírez-Esparza, N., Slatcher, R.B., Pennebaker, J.W.: Are women really more talkative than men? Science 317(5834), 82–82 (2007), http://science.sciencemag.org/content/317/5834/82
18. Podesva, R., D'Onofrio, A., Van Hofwegen, J., Kim, S.: Country ideology and the California Vowel Shift. Language Variation and Change 27(2), 157–186 (2015)
19. Precht, K.: Sex similarities and differences in stance in informal american conversation. Journal of Sociolinguistics 12(1), 89–111 (2008), https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9841.2008.00354.x
20. Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
21. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter (Sep 2017)
22. Sadker, D., Sadker, M.: Is the O.K. classroom O.K.? The Phi Delta Kappan 66(5), 358–361 (1985), http://www.jstor.org/stable/20387346
23. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
24. Schütze, H.: Word space. In: Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93. pp. 895–902. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
25. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)