New Machine Learning Methods Demonstrate the Existence of a Human Stylome

Hans van Halteren[*], R. Harald Baayen[†], Fiona Tweedie[‡], Marco Haverkort[§] & Anneke Neijt

[*]*Department of Language and Speech, University of Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands*

[†]*Max-Planck-Institut für Psycholinguistik, P.O. Box 310, NL-6500 AH, Nijmegen, The Netherlands*

[‡]*School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh EH9 3JZ, UK*

[§]*Department of Linguistics, University of Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands (also Department of Linguistics, Boston University)*

Department of Dutch, University of Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands

Corresponding author:
Hans van Halteren
Email hvh@let.kun.nl (preferred)
Telephone +31 26 361 2836

Abbreviated title:
The Existence of a Human Stylome

New Machine Learning Methods Demonstrate the Existence of a Human Stylome

Abstract

Earlier research has shown that established authors can be distinguished by measuring specific properties of their writings, their stylome as it were. Here, we examine writings of less experienced authors. We succeed in distinguishing between these authors with a very high probability, which implies that a stylome exists even in the general population. However, the number of traits needed for so successful a distinction is an order of magnitude larger than assumed so far. Furthermore, traits referring to syntactic patterns prove less distinctive than traits referring to vocabulary, but much more distinctive than expected on the basis of current generativist theories of language learning.

We all make extensive use of a natural language to communicate with others. The ease with which we do this might give the impression that all the users of a particular natural language are using the exact same language. Now, specific languages do not come completely hard-wired in the brain, although the brain is thought to contain a set of hard-wired expectations about the structure of natural language, the so-called Universal Grammar. Whether or not there is indeed some support from structures in the brain, we always have to learn a

language on the basis of examples of its use. However, the set of examples we are exposed to while learning a language, especially our native language, is pretty much unique for each of us. We therefore necessarily create our own unique form of the language, which merely appears to be an instantiation of 'the same language' because the form is derived from a sufficiently similar set of examples. This much is generally taken to be true by most linguists (Atkinson, 1992; Chomsky, 1999; Gopnik, 1997; Guasti, 2002; Harris, 1998; Jenkins, 2000; O'Grady, 1999; Pinker, 2002; Rice, 1996; Strozer, 1994; Wexler, 1999; for a different perspective, see Elman et al., 1996; Karmiloff & Karmiloff-Smith, 2002; Tomasello, 1999). What is much less clear, however, is how much these individual language forms differ, and whether the differences are systematic and measurable (Schütze, 1996). In other words, it is unclear if individual language forms can be classified in terms of a "stylome", a set of measurable traits of language products. Here we will attempt to identify such a stylome, more specifically a stylome which is extensive enough to be able to distinguish between pairs of language users on the basis of their language use. Another question is whether the differences between individual language forms are equally visible for all aspects of a language. In the generative tradition, most theories about language learning predict that the syntactic structures of our native language are reasonably fixed once we reach the age of eight (at the latest), but that vocabulary keeps

growing throughout our lives (Clark, 1993). If the grammar indeed stabilizes early, vocabulary is the only place where the language can adapt to changes in the conditions under which we use the language. This implies that the most useful components of a stylome would refer to word use, while the use of syntactic patterns should be much less distinctive. If we manage to identify a stylome of sufficient quality, our secondary question will therefore be whether the predicted dichotomy between vocabulary and syntax is indeed visible in the relative usefulness of stylome components.

Experimental Task

The first steps in our investigation into the existence of a stylome are the identification of a) a useful set of measurable traits and b) a benchmark task to prove their value. For both we can look initially to the field of authorship attribution. Its practitioners are most often concerned with scholarly subjects like general history (Mosteller & Wallace, 1984) and literary history (Holmes, 1998), but may also work on more practical problems like criminal investigations (Broeders, 2001; Chaski, 2001). In authorship attribution circles we find a number of traits which are used as a matter of tradition, like vocabulary richness, or the word counts of the 50 most frequent function words (Burrows, 1992). We currently also see an active drive for the discovery of more traits, as well as the desire for a stringent examination of some accepted fundamental truths of the field (Rudman, 1998; Grant & Baker,

2001). One of the reasons for this activity is a recent study on within-text comparison, using crime fiction by two different authors (Baayen et al., 1996). This study shows that, although the two authors differ very clearly (and measurably) in style, the traditional methods fail occasionally to attribute text samples correctly. This shows that even though traditional traits may well suffice for the identification of most established authors, possibly because such authors have developed a recognizable personal style on purpose, a more generally useful stylome will need to encompass more, perhaps much more. Another lesson from the study is that distinguishability is a matter of degree: some authors are easier to recognize than others. Many of the existing benchmark texts are probably of the easier type, since they could be handled with the traditional methods. Many others cannot be used either as their true authors are not actually known. If we really want convincing evidence for the existence of a general measurable stylome, we need to create our own benchmark task, and especially one which is designed to be hard.

With the goal of a hard benchmark task in mind, we set out to compile a Dutch Authorship Benchmark Corpus (ABC-NL), starting with a component which focuses on widely divergent written texts produced by very similar authors (ABC-NL1). The ABC-NL1 corpus consists of 72 Dutch texts by 8 authors, controlled for age and educational level of the authors, and for register, genre and topic of the texts. The authors were students of Dutch at the

University of Nijmegen. We selected these specific authors with the expectation that their language skills would be advanced, but their writing styles would as yet be at most weakly developed and hence very similar, unlike those of the authors in standard attribution problems. Each author was asked to write nine texts of about a page and a half. In the end, it turned out that some authors were more productive than others, and that the text lengths varied from 628 to 1342 words. The authors did not know that the texts were to be used for authorship attribution studies, but instead assumed that their writing skill was measured. The topics for the nine texts were fixed, so that each author produced three argumentative non-fiction texts, on the television program Big Brother (a1), the unification of Europe (a2) and smoking (a3), three descriptive non-fiction texts, about soccer (d1), the upcoming new millennium (d2) and the most recent book they read (d3), and three fiction texts, namely a fairy tale about Little Red Riding Hood (f1), a murder at the university (f2) and a chivalry romance (f3).

To verify that every pair of authors can indeed be distinguished on the basis of a proposed stylome, we have to execute a 2-way classification task. For each author (8), each topic (9, times 8 authors leads to 72 texts) and each alternative author (7, times 72 texts leads to 504 trials), we create a classification model which is based on the texts written by the two authors on the other eight topics. We then check whether this model assigns the text

under investigation to the actual author of the text. If personal language use is not systematic/measurable, classification should be random and correct classification ought to occur in 50% of the trials. The higher the actual score of the classification procedures, the stronger the case for a measurable stylome.

The hardness of the ABC-NL1 classification task is demonstrated if we try to perform it with traditional methods. Using the overall relative frequencies of the fifty most frequent function words and a Principal Components Analysis (PCA) on the correlation matrix of the corresponding 50-dimensional vectors shows no discernable authorial structure at all (Baayen et al., 2002). The use of Linear Discriminant Analysis (LDA) on overall frequency vectors for the 50 most frequent words leads to classification scores of around 60 percent, which can be increased to around 80 percent by using cross-sample entropy weighting (Baayen et al., 2002). These scores do show that a vocabulary-related stylome has potential. However, if we want to firmly establish the existence of a stylome, we need our classification scores to be much nearer to 100 percent.

Approach and Results

To obtain these more convincing classification scores, we have developed a new classification approach, in which we do not use a single summary vector of the contents of each text, but rather a large number of local observations. For every token (word or punctuation mark) in the text, we determine the

following properties:

1. Current token
2. Previous token
3. Next token
4. Concatenation of the wordclass tags of these three tokens (as assigned by an automatic WOTAN-lite tagger; van Halteren et al.., 2001)
5. Concatenation of
    a. length of the sentence (in 7 classes: 1, 2, 3, 4, 5-10, 11-20 or 21+ tokens)
    b. position in the sentence (in 3 classes: first three tokens, last three tokens, other)
6. Concatenation of
    a. part of speech of the current token, i.e. the initial part of the wordclass tag
    b. frequency of the current token in the text (in 5 classes: 1, 2-5, 6-10, 11-20 or 21+)
    c. number of blocks (consisting of 1/7th of the text) in which the current token is found (in 4 classes: 1, 2-3, 4-6,

7)

d. distance in sentences to the previous occurrence of the current token (in 7 classes: NONE, SAME, 1, 2-3, 4-7, 8-15, 16+)

We then combine these six properties into a six-dimensional feature vector.

We use the Weighted Probability Distribution Voting algorithm (WPDV; see Appendix I) to derive a classification model from the observed feature vector sets. As the algorithm is sensitive to text length (by way of the observation set size), we actually use only 700 observations for each text (randomly selected) for a specific model. Each WPDV model is therefore trained on a collection of 11200 (2 authors x 8 training texts x 700 observations) feature vectors, each with an indication of the corresponding author. As the WPDV system considers all combinations of features within the vectors as traits as well, the number of traits is much larger than in the traditional methods, viz. around 500,000 occurring feature combinations. After training, the model is applied to each of the observed feature vectors for the test text, leading to a large number of probability estimates for each candidate author A_i, which are then translated (see Appendix II) into a single overall estimate $P(A_i)$.

[INSERT FIGURE 1 AROUND HERE]

The probability with which texts are assigned to the correct author is shown in Figure 1. Of the 504 trials, 493 are successful (97.8%). Furthermore, for erroneous choices both $P(A_i)$ are generally fairly close to 0.5. The most extreme wrong prediction is the assignment of text f2 by author 8 to author 4, with a probability of 0.550 versus 0.450 for author 8 himself. This means that, if we were to plot a precision/recall curve for the various possible thresholds, 100% precision would be reached at threshold 0.551. The recall at this point is 90.1% (454 trials). In other words, we could set a (post-hoc) confidence threshold of 0.551, and only let the system report attributions if $P(A_i)$ is higher than the threshold. Under those circumstances the system would only suggest an author for 454 of the trials (90.1%), but these would all be correct.

[INSERT FIGURE 2 AROUND HERE]

The successes and failures in attribution are distributed regularly across the various authors and the various text types, as shown in Figure 2. Although one might have expected there to be “easy” and “hard” authors, it appears that each is well-recognizable for most texts. Still, each author is occasionally unrecognized or falsely recognized. The same situation holds for the various text types and classes.

Vocabulary versus Syntax

These results are such that we no longer need to doubt the general existence of a measurable stylome, and can advance to our second question: the relative strengths of vocabulary-related and grammar-related traits. Ideally, a rich notion of syntax, which includes hierarchical relations, co-reference, long distance dependencies, etc., should be used. For computational reasons, however, we are forced to refer to surface phenomena only, viz. the word class tag concatenation mentioned above. Even so, surface variation is still constrained to such an extent by more abstract grammatical principles that it should allow for relatively little variation. We examine the relative strengths of vocabulary and grammar by selecting subsets of the six features used above. The first three features (the three actual tokens) are used to create a model using only vocabulary-related traits, which we call VOCAB. The fourth feature, consisting of the wordclass tags for the three tokens, is used to create a model referring to syntactic usage, which we call SYNT. However, the feature is split out into three separate features, viz. the three individual tags, e.g. f_{tags}="Prep/Pron(aanw,neut)/N(ev,neut)" becomes f_{prev}="Prep", f_{cur}="Pron(aanw,neut)" and f_{next}="N(ev,neut)". As can be expected, the SYNT model has access to less input information, viz. about 8,000 occurring feature combinations versus about 40,000 for the VOCAB model. This quantitative difference alone would suggest that SYNT ought to produce a lower correct than VOCAB, but the

linguistic musings above would predict that there is also a qualitative difference, which should prevent the SYNT model from reaching any high correct attribution percentage at all.

In the actual test, we compare the two models on the already familiar score, the percentage of the texts which are assigned to the correct author, which is 97.8% for the model above. However, we do not just measure these scores for full texts, but also for shorter stretches of text. After all, the switch from the traditional summary vectors to collections of local observations has the additional advantage that there is no longer any statistically inspired lower limit on the size of the test text. We did not use this potential in the model above, because some of the traits themselves, e.g. token distribution, refer to a wider context. But the traits used by VOCAB and SYNT are purely local. This means that the number of available observations on which the system bases its classification can be varied freely, down to even a single observation. At the shorter stretch lengths we use several different stretches in order to compensate for a greater expected variation in measurements, e.g. for the experiments with 10 observations, we use 60 stretches per test text (x 504 comparisons = 30240 trials).

[INSERT FIGURE 3 AROUND HERE]

The results of the new comparisons are shown in Figure 3. When used on complete texts, VOCAB is only slightly worse (97.0%) than the full model above (97.8%). At lower numbers of observations, i.e. shorter text stretches, the attribution scores are of course lower, but they stay surprisingly high. Even for extremely low numbers of observations, the scores are significantly higher than chance. At all levels, SYNT perform worse than VOCAB, but certainly not as much worse as expected.

Conclusions

From our experiments we draw several conclusions. First of all, it is obvious that we have succeeded in identifying measurable traits which are characteristic of our eight authors' language use. The models used here may not be able to attribute texts with 100% certainty, but the scores are clearly significantly higher than can be explained without reference to such measurable traits. Also, the extreme effectiveness of the models shows that the differences between the "personal" language versions of even non-specialist writers are greater than expected so far.

Vocabulary traits are generally more useful than syntactic traits, as predicted, but the difference is much less pronounced than expected. Even when restricted to purely syntactic traits the system still produces a correct attribution in 88.7% of the trials (using complete texts). Given these results, it

becomes much less likely that it is true that there is a clear division between grammar development, with stabilization at an early age, and vocabulary development, without stabilization.

Finally, the quality of the attribution increases when the number of traits that is taken into account increases. This effect is clearly visible in the relation between number of observations and the resulting classification quality. The effect is also present in the relative quality of the three examined models. This implies that even better attribution might be possible if even more traits are included in the stylome, and that it is worthwhile to embark upon a systematic search for further useful traits (Rudman, 2000). It is unlikely that we can ever reach a 100% correct attribution for every specific writer, especially when the writer is aware of our attempts and is consciously trying to manipulate his writing style (Pawlowski, 1998). However, our experiments do lead us to believe that we can get very close, even though this may well force us to use a stylome which is yet another order of magnitude larger.

References

Atkinson, M. (1992). *Children's Syntax*. Cambridge, Mass.: Blackwell.

Baayen, R.H., van Halteren, H. and Tweedie, F. (1996). "Outside the Cave of Shadows: Using syntactic annotation to enhance authorship attribution". *Literary and Linguistic Computing* **7**: 91-109.

Baayen, R.H., van Halteren, H., Neijt, A., and Tweedie, F. (2002). "An Experiment in Authorship Attribution". Proceedings JADT 2002, pp. 69-75.

Broeders, A.P.A. (2001). "Forensic Speech and Audio Analysis, Forensic Linguistics 1998-2001 – A Review". Proceedings 13th Interpol Forensic Science Symposium, Lyon, France.

Burrows, J.F. (1992). "Computers and the Study of Literature". In: Butler, C.S. (ed.). *Computers and Written Texts.* Oxford: Blackwell, pp. 167-204.

Chaski, C.E. (2001). "Empirical Evaluations of Language-Based Author Identification Techniques". *Forensic Linguistics* **8(1):** 1-65.

Chomsky, N. (1999). "On the Nature, Use, and Acquisition of Language". In: Ritchie, W. & Bhatia, T. (eds.). *Handbook of Child Language Acquisition.* San Diego: Academic Press, pp. 33-54.

Clark, E.V. (1993). *The lexicon in acquisition.* Cambridge: Cambridge University Press.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, Mass.: MIT Press.

Gopnik, M. (ed.). (1997). *The Inheritance and Innateness of Grammars*. Vancouver Studies in Cognitive Science, Volume 6. Oxford: Oxford University Press.

Grant, T., and Baker, K. (2001). "Identifying Reliable, Valid Markers of Authorship: a response to Chaski". *Forensic Linguistics* 8(1) 66-79.

Guasti, M.-T. (2002). *Language Acquisition.* Cambridge, Mass.: MIT Press.

van Halteren, H., Zavrel, J. and Daelemans, W. (2001). "Improving accuracy in word class tagging through the combination of machine learning systems". *Computational Linguistics* **27(2):**199-230.

van Halteren, H. (2001). "A default first order weight determination procedure for WPDV models". Proceedings CoNLL 2001, pp. 119-122.

Harris, J. (1998). *The Nurture Assumption.* New York: The Free Press.

Holmes, D.I. (1998). "Authorship attribution". *Literary and Linguistic Computing* **13(3):**111-117.

Jenkins, L. (2000). *Biolinguistics: Exploring the Biology of Language.* Cambridge: Cambridge University Press.

Karmiloff, K. & Karmiloff-Smith A. (2001). *Pathways to Language: From*

Fetus to Adolescent. Cambridge, Mass.: Harvard University Press.

Mosteller, F., and Wallace, D.L. (1984). *Applied Bayesian and Classical Inference in the Case of the Federalist Papers* (2nd edition). Springer Verlag, New York.

O'Grady, W. (1999). "The Acquisition of Syntactic Representations: A General Nativist Approach". In: Ritchie, W. & T. Bhatia (eds.). *Handbook of Child Language Acquisition.* San Diego: Academic Press, pp. 157-194.

Pawlowski, A. (1998). *Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Emile Ajar.* Paris, Genève: Champion-Slatkine.

Pinker, S. (2002). *The Blank Slate: The Modern Denial of Human Nature.* London: Allen Lane.

Quinlan, J.R. (1986). "Induction of Decision Trees". *Machine Learning* **1**:81-206.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman, San Mateo, CA.

Rice, M. (ed.). (1996). *Towards a Genetics of Language.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rudman, J. (1998). "The state of the authorship attribution studies: some problems and solutions". *Computers and the Humanities* **31**:351-365.

Rudman, J. (2000). “The style-marker mapping project: a rationale and progress report”. ALLC/ACH 2000 Conference Abstracts.

Schütze, C. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology.* Chicago: The University of Chicago Press.

Strozer, J. (1994). *Language Acquisition after Puberty.* Washington D.C.: Georgetown University Press.

Tomasello, M. (1999). *The Cultural Origin of Human Cognition.* Cambridge, Mass.: Harvard University Press.

Wexler, K. (1999). “Maturation and Growth of Language”. In: Ritchie, W. & Bhatia, T. (eds.). *Handbook of Child Language Acquisition.*San Diego: Academic Press, pp. 55-110.

Appendix I: Weighted Probability Distribution Voting

Weighted Probability Distribution Voting (WPDV; van Halteren, 2001) is a supervised learning approach to the automatic classification of items. The set of information elements about the item to be classified, generally called a "case", is represented as a set of feature-value pairs, e.g. the set F_{case} = { f_{prev}="in", f_{cur}="dit", f_{next}="gebied", f_{tags}="Prep / Pron(aanw,neut) / N(ev,neut)", f_{sen}="21+ / mid", f_{distr}="Pron / 2-5 / 2-3 / NONE" } from the authorship attribution task. The values are always treated as symbolic and atomic, not e.g. numerical or structured, and taken from a finite (although possibly very large) set of possible values. An estimation of the probability of a specific class for the case in question is then based on the number of times that class was observed with those same feature-value pair sets in the training data. To be exact, the probability that class C should be assigned to F_{case} is estimated as a weighted sum over all possible subsets F_{sub} of F_{case}:

$$P(C) = N(C) \sum_{F_{sub} \subseteq F_{case}} W_{F_{sub}} (freq(C \mid F_{sub}) / freq(F_{sub}))$$

with the frequencies (freq) measured on training data, and N(C) a normalizing factor such that $\sum_C P(C) = 1$.

In principle, the weight factors $W\{F_{sub}\}$ can be assigned per individual subset. For the time being, however, they are assigned per 'family' of subsets,

e.g. the family of all the subsets consisting of the features f_{cur} and f_{next}. In the current experiments, the weight for each family is set to the product of the family's component features' Gain Ratio values (Gain Ratio being a normalised derivative of Information Gain; Quinlan, 1986 & 1993) times an optimal multiplication.

Appendix II: Combination of local probability estimates to create overall estimate.

During the attribution of texts, a large number of probability estimates $P_j(A_i)$ for each candidate author $A_{i\ has}$ has to be translated into a single overall estimate $P(A_i)$. This is done with a thresholded and weighted addition:

$$P(A_i) = C_N \sum_{j\ \ observations} IF\ P_j(A_i) > 0.5\ THEN\ (P_j(A_i) - 0.5)^D\ ELSE\ 0$$

in which C_N is a constant to normalize the totals so that $P(A_1) + P(A_2) = 1$. The parameter D in the addition can be used to give more weight to more decisive local estimates. The higher the value of D, the more dominating the extremely confident local estimates are in the overall estimate. The best performance for ABC-NL1 is reached at D=3.

To further improve the quality of the process, we execute each training-application run three times, each time using a different set of 700 selected observations. In this way we arrive at three overall estimates for the text,

which we then combine (van Halteren et al., 2001; here by simple averaging) to determine an ultimate decision on the authorship of the text.

FIGURE 1

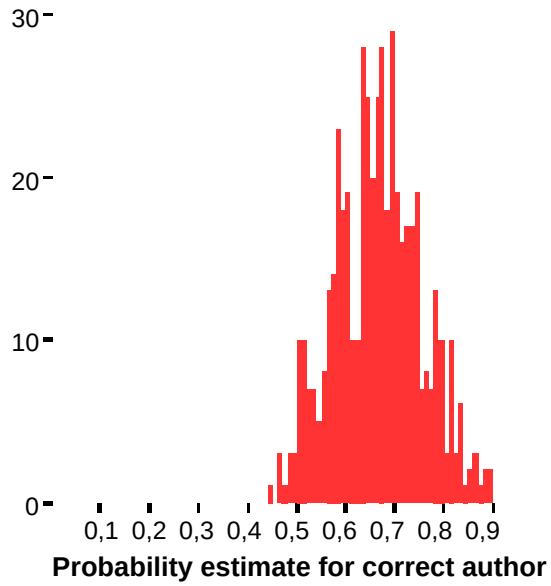
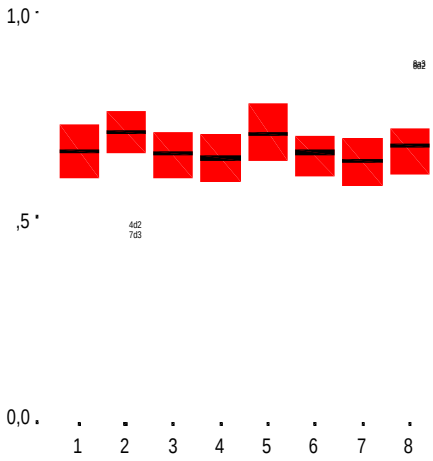

30
20
10
0
0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9
Probability estimate for correct author

FIGURE 2



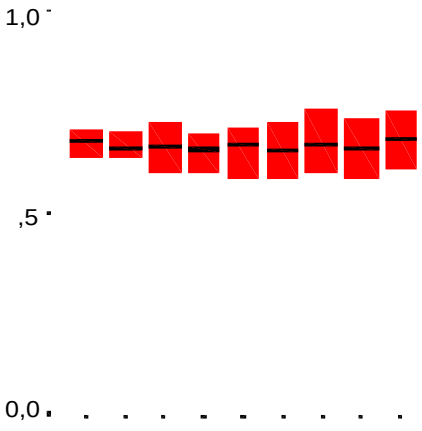
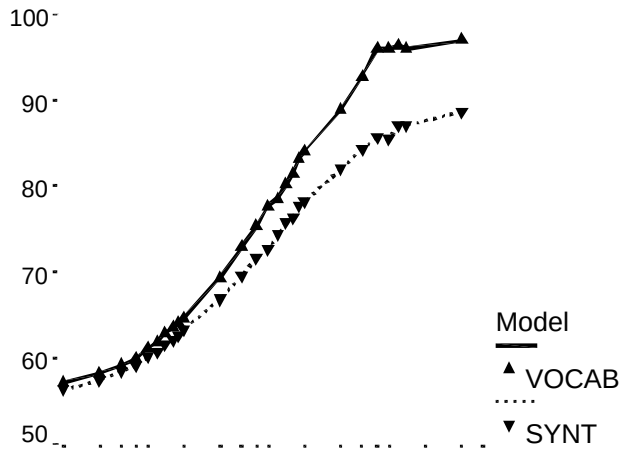

1,0
,5
0,0
Text subject id

FIGURE 3



Number of observations (size of test text)

FIGURE LEGENDS

Figure 3. A histogram for the probability with which texts are assigned to the correct author. Texts with scores above 0.5 are assigned correctly. Most texts are assigned with reasonable confidence and also correctly. The 11 erroneous attributions are all found in the lower confidence range.

Figure 3. Box plots for the probability with which texts are assigned to the correct author, grouped per author participating in the comparison (plot on left) and per text subject (plot on right).

Figure 3. Classification scores for the VOCAB and SYNT models as a function of the number of observations available to the system, i.e. the size of the test text.