

# Authors, Genre, and Linguistic Convention

Jussi Karlgren and Gunnar Eriksson  
Swedish Institute of Computer Science  
jussi@sics.se, guer@sics.se

## Authorship, Language, and Individual Choice

The basic premise underlying authorship attribution studies is that while the form of expression in language in some respects is strictly bound by linguistic rule systems, in others somewhat constrained by topic and genre, it is in some other respects freely available for configuration or preferential choice by author or speaker. This individual variation can be observed, detected, and predicted to some extent, using traditional stylostatic measures: e.g. word length varies from author to author [Mendenhall, 1887, e.g.]; sentence length likewise; some forms of lexical expression are characteristic of speakers, either on an individual level or on community level [Book of Judges].

Common to most computation of individual difference in authorship is that the features used to characterise and distinguish authors are *local*, based on the repeated computation of some statistic at various positions in the text and then averaging or normalising the result. In this position paper we claim that local features are subject to pressure from conventionalisation and grammaticalisation processes in language, and that textually *global* features should be better suited for the distinctions we are after: individual choice of informational organisation.

## Rules, Constraints, and Conventions

The patent regularities of linguistic expression are formed by constraints – rules, conventions, or norms of e.g. biological, social, psychological, or communicative, character. While language use is regular to a great extent, the rules that govern it change continuously. Observations and descriptions of language from an earlier time can become obsolete; early samples of language can be all but incomprehensible to the modern reader (and presumably, to the listener). The origin of linguistic constraints, their ontological nature, and their life span or life cycle is much debated in linguistics, but grammaticalisation, the process whereby optional linguistic behaviour becomes a norm, is assumed to proceed sequentially, with many partially counteracting motivating

factors and driving forces, variously ascribed to economy of expression, redundancy, tolerance towards noise, and factors related to social cohesion [Dahl, 2006, e.g.].

Many obligatory grammatical features are likely to have started their life as idiosyncratic choice, become accepted in some community as marker of some function, informational or social, and thence migrated to broader linguistic usage.

Given this underlying process from characteristics of individual usage to conventionalisation, and further to grammatical constraints, the claim underlying these first experiments is that the degree of leeway or freedom for the individual user varies, not only between some specific features, but between some *types* of observable features: text-global patterns, e.g. being less rule-bound than local clause-internal structure.

In brief, features grammaticalise, but first conventionalise. Some features are optional, some non-optional. All features are not as accessible to the process of grammaticalisation. The features most studied in the fields of linguistics, information access, and stylostatics are lexical or syntactic on a local level. These are precisely the situation-independent, topic-independent, speaker-independent features most susceptible to linguistic control and grammaticalisation (for the purposes of this discussion, folding in lexicalisation as a special case of conventionalisation).

There is good reason for syntacticians to study the local and rule-bound variation – the task of linguistics is to generalise from observations to rules; for the purposes of authorship attribution the converse is the case – the task is to find the characteristic and the special. Global textual patterns are available for author choice and will provide better purchase for discrimination of individual style than choice on a level where conventions are strong, observable usage for language users less sparse, and grammar and grammaticalisation hold fast.

Free	Author	Repetition, organisation, elaboration
Convention	Genre	Lexical patterns, patterns of argumentation, tropes
Rule	Language	Syntax, morphology

Figure 1: Levels of constraints.

## Observanda — Features

What sort of features do we, as authorship attribution experimentalists, then have recourse to? Typically text categorisation studies compute observed frequencies of some lexical items, or some identifiable construction. An observed divergence in a text sample from expected occurrence of that specific item (with prior information of taken into account) is a mark of individuality and can be used in the process of identifying author (or, indeed, similarly, genre, or topic).

This type of variation, however, is heavily bound by local textual structure and not as likely to yield as much individual variation as will variation as measured over the length of the text, on the level of information organisation: examples might be term recurrence [Katz, 1996] or term patterns [Sarkar, 2005]; type token ratio [Tallentire, 1973]; rhetorical structure; measures of cohesion and coherence [Halliday, 1978]; measures of lexical vagueness, inspecificity, and discourse anchoring; and many other features with considerable theoretical promise but rather daunting computational requirements.

In the present first experiment two simple binary features were calculated:

**adv** the occurrence of more than one adverbial of any type in a sentence, and

**clause** the occurrence of more than two clauses of any type in a sentence.

Each sentence was thus given the score 1 or 0 for each of the two features. The choice of features was purposely kept simple – both these features are simple to compute, but have pertinence to informational and topical organisation: “clause” being a somewhat more sophisticated proxy for syntactic complexity than sentence length; “adv” a measure of topical elaboration.

## Corpus

For this experiment, one year of newsprint from the Glasgow Herald was used, about 34 000 articles in all. About half of the articles are tagged for “Article type”, and 28 000 have a byline. 8 article types, as given in Table 2, are found in the collection, and 244 authors with more than 500 sentences. The texts were preprocessed by the Connexor tool kit for English morphology, surface syntax, and syntactic dependencies.

## Measurements and metrics

The measurements for the two chosen variables are given in Table 3. The table shows, somewhat unsurprisingly, that the palette of genres is more consistent with each other than are authors (the number of authors is large; only the authors with the highest and lowest scores for each feature are shown): some authors have really very few clauses and very few adverbials in their sentences, but all genres have a somewhat consistent density of subclauses and adverbials.

## Transition patterns

Returning to the main claim of this paper, we investigate whether the introduction of longitudinal features spanning

<ARTICLETYPE>	<i>n</i>
advertising	522
book	585
correspondence	3659
feature	8867
leader	681
obituary	420
profile	854
review	1879
<b>total tagged</b>	<b>17467</b>

Figure 2: Sub-genres of the Glasgow Herald.

	feature	clause	adv
advertising	0.899	0.682	
book	0.832	0.637	
correspondence	0.918	0.705	
feature	0.929	0.689	
leader	0.931	0.735	
obituary	0.784	0.601	
profile	0.921	0.712	
review	0.866	0.646	
author $c_{max}$	0.96		
author $c_{min}$	0.52		
author $a_{max}$			0.81
author $a_{min}$			0.39

Figure 3: Relative presence of features “clause” and “adv” in sentences

over text rather than local measurements might improve the potential for categorisation of authors. To do this, the two features studied were measured over a rolling window of one to five sentences, and the resulting transition pattern was recorded over each text. The first and last bits of text where the window length would have extended over the text boundary was discarded. For windows of size two, four possible patterns were tabulated, for windows of size five, thirty-two, as shown in Figure 4. The experiment is designed to investigate whether using such longitudinal patterns improves the potential for author identification *more* than it improves the potential for genre identification.

window size	patterns	number patterns
1	1, 0	2
2	11, 10, 01, 00	4
3	111, 110, 101, 100 011, 010, 001, 000	8
4	1111, ..., 0000	16
5	11111, ..., 11101, 11100, ..., ..., 00000	32

Figure 4: Feature space for varying window size

## Models of probability

The observed presence of a feature in a pattern, normalised for sentence frequency, yield a crude estimate of probability of recurrence of any observed pattern in further texts in the

same category – the same genre or same author. Such a probability distribution describes the density of occurrence over the different features values – how often some feature is likely to occur, compared to the others.

Thus, as an example, any text in category “correspondence”, using a feature space based on a window size of three, has the relative probabilities describable as a vector of probability estimates – and is likely to have about two thirds of sentences in runs without multiple clauses:

$$\begin{aligned}
 p_3(\text{correspondence}) &= \\
 &= \{p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000}\} = \\
 &= \{0.0069, 0.0654, 0.00903, 0.155, 0.00454, 0.0363, 0.0486, 0.674\}
 \end{aligned}$$

## Evaluation

In all categorisation tasks, an unknown item – text, in this case – with an observation or estimate of feature values, is matched to the category closest to it – in some way, using some algorithm, and some definition of “closest”. In these experiments we choose not to test our probability distributions applied to categorisation, to avoid the noise necessarily introduced by the categorisation methodology itself, but instead use an intrinsic assessment of the probability distributions over the target categories.

The assumption we make is that if the set of target categories is well distributed over the feature space, matching unknown items to it will be easier than if they are clustered together. Or, in other words, one wishes to find features which separate categories well. So, given a feature space we wish to use some measure for how widely it separates the target categories at hand. Figure 5 shows the probability estimates for the eight genres and some randomly picked authors in the material with a window size of 2 for the feature “clause”. The question is how distinct this estimate finds the categories to be.

Distance between probability distributions are commonly measured or assessed using the Kullback-Leibler divergence measure [Kullback and Leibler, 1951]. Since the measure as defined by Kullback and Leibler is asymmetric, we use a symmetrised version, a harmonic mean given by [Johnson and Sinanović, 2001].

$$d_{kls} = \frac{1}{\frac{1}{\sum_{i=0}^n p_i \times \log_2(p_i/q_i)} + \frac{1}{\sum_{i=0}^n q_i \times \log_2(q_i/p_i)}}$$

The divergence is a measure of distance between two categories. In this experiment, for each condition, we report a sum of pairwise divergences for the set of categories.<sup>1</sup> A large figure indicates a greater separation between categories – which is desirable from the perspective of a categorisation task, since that would indicate better potential power for

<sup>1</sup>This is not necessarily the correct or most informative way of using this measure. We might e.g. instead compute and report the divergence between the two closest categories.

Window size	Genre		Author	
	“clause”	“adv”	“clause”	“adv”
1	0.5129	0.1816	0.7254	0.4033
2	0.8061	0.3061	1.3288	0.8083
3	1.1600	0.4461	2.1577	1.2211
4	1.4556	0.6067	2.3413	1.8111
5	1.7051	0.7650	3.0028	2.2253

**Figure 6: Occurrence patterns’ effect on features “clause” and “adv”**

working as a discriminating measure between the categories under consideration.

The category set is vastly different for authors and genres. There are eight genres and 244 authors with more than 500 sentences. The sums of pairwise divergences for the two category sets are of different orders of magnitude, and in order to facilitate comparison between authors and genres, we randomly select eight authors, compute the sum of pairwise differences for that set, repeat this fifty times (with replacement), and use the mean of the resulting divergences as the result.

For each window length, the sum of the symmetrised Kullback-Leibler measure for all genres or authors is shown in Figure 6. The figures can only be compared line by line in the table – the divergence figures for different numbers of features, as is the case for different window sizes, cannot directly be related to each other. This means that we cannot directly say if window size improves the resulting representation or not, in spite of the larger divergence values for larger window size. Given that caveat, the relative difference between the features can be compared, and the table gives us purchase to make two claims.

Firstly, comparing both features for each window size between genre and author representations we find that the *difference* between genre categories and author categories is greater for large window sizes. This speaks to the possibility of our main hypothesis holding: larger window size allows a better model of individual choice than a shorter one.

Secondly, we find that feature “adv” seems to make relative gains compared to feature “clause” for the larger window size, for the author case: while “clause” still shows a larger value, the relative difference is smaller for the larger window size. This speaks to the possibility of finding possibly better informed feature sets for the larger contextual models to improve distinction between individuals rather than genres.

genre	$p_{11}$	$p_{10}$	$p_{01}$	$p_{00}$
feature	0.022	0.078	0.056	0.84
review	0.041	0.13	0.072	0.76
advertising	0.011	0.072	0.039	0.88
profile	0.016	0.056	0.040	0.89
leader	0.016	0.055	0.023	0.91
correspondence	0.066	0.15	0.051	0.73
obituary	0.0079	0.072	0.023	0.90
book	0.038	0.084	0.069	0.81
author	$p_{11}$	$p_{10}$	$p_{01}$	$p_{00}$
Stephen McGinty	0.013	0.071	0.052	0.86
Ian Paul	0.021	0.050	0.018	0.92
James O'Brien	0.018	0.11	0.088	0.78
Hugh Dan MacLennan	0.19	0.097	0.032	0.68
Tom McConnell	0.013	0.11	0.052	0.82
William Tinning	0.0062	0.071	0.020	0.90
Andrew Mackay	0.018	0.063	0.038	0.88
Charlie Allan	0.0067	0.047	0.032	0.91
Robert Ross	0.010	0.064	0.027	0.90

Figure 5: Probability estimates for genres and some authors, window size 2, feature “clause”

## Conclusions

This experiment was a first shot at finding whether more sequential features might not be better than local ones for distinguishing between genres and authors.

Our *topical goal*, for these experiments, restated, is that lengthier text spans might give better purchase for finding features open for author choice as compared to locally computed features, mostly determined by syntax. Adverbials, as an example, might be expected to have a certain occurrence frequency in any genre or topic, but the placement of them in text and their consequent distribution can be assumed to be up to individual choice rather than genre or topical convention or syntactic constraint.

The results cannot be said to give conclusive evidence to show whether our hypothesis holds or not, but they do encourage further study.

At this juncture the task is finding more (and more informative) features and factors of the more non-conventional levels of the linguistic system, measuring them, evaluating them, and understanding and diagnosing their import on the knowledge representation we choose for the application we intend to work with. The features we expect to study are intended to reach beyond occurrence statistics, measure presence or repetition rather than frequency, avoid notions such as average and mean and instead model patterns, trends, bursts and variation.

The *methodological goal* of the experiment is to build an experimental process based on hypotheses informed by some sense of textual reality, rather than computational expediency, and to evaluate the results by discriminatory power, not by application to noisy task. We will further investigate the diagnostic power of e.g. divergence measures, rather than sample-based experiments, to study the potential usefulness of a knowledge representation scheme.

## Choice points left by the wayside

Some questions clamor for attention in this specific experimental setting:

- Is Kullback Leibler divergence (and its current symmetric implementation) the right measure to measure distance between observed occurrence patterns?
- Is summing pairwise divergences the best way of modelling the consistency of a set of category models? Maybe measuring the separation between the two closest neighbours in a set would be better?
- If we would happen to be convinced that transitional patterns are better than local singularities as a feature base – is the model presented here a step in the right direction?
- What better kernel features – beyond adverbial and clause count – should we try utilising?

## Acknowledgments

This experiment was funded by the Swedish Research Council. We are grateful to our colleague Anders Holst for providing us with formulæ and valuable intuitions and starting points.

## 1. REFERENCES

In *Book of Judges, King James Version, Old Testament*, chapter 12, pp. 5–6.

Östen Dahl. 2006. *The Growth and Maintenance of Linguistic Complexity*. John Benjamins, Amsterdam, Philadelphia.

M A K Halliday. 1978. *Language as social semiotic*. Edward Arnold Ltd, London.

Don H Johnson and Sinan Sinanović. 2001. “Symmetrizing the Kullback-Leibler distance”. *IEEE Transactions on Information Theory*.

Slava Katz. 1996. “Distribution of content words and phrases in text and language modelling”. *Natural Language Engineering*, 2:15–60.

S Kullback and R A Leibler. 1951. “On information and sufficiency”. *Annals of Mathematical Statistics*, 22:79–86.

T.C. Mendenhall. 1887. “The Characteristic Curves of Composition”. *Science*, 9:237–249.

Avik Sarkar, A de Roeck, and P H Garthwaithe. 2005. “Term re-occurrence measures for analyzing style”. In *Textual Stylistics in Information Access. Papers from the workshop held in conjunction with the 28th International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, August. ACM SIGIR.

D. Tallentire. 1973. “Towards an Archive of Lexical Norms: A Proposal”. In A. Aitken, R. Bailey, and N Hamilton-Smith, editors, *The Computer and Literary Studies*. Edinburgh University Press.