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 is in some respects strictly bound by linguistic rule systems and 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 stylostatistic measures. For instance, word length varies from author to author [Mendenhall, 1887, e.g.]; sentence length likewise; and some forms of lexical expression are characteristic of speakers, either on an individual level or on a 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 based on the repeated measurement of some, often clause-internal, property at independent positions in the text and then *aggregating* these *pointwise* measures by averaging or normalising the result. In this position paper we claim that by measuring *local* clause- or even word-internal properties, and by aggregating in such a way that the relation between individual observations is destroyed, we obtain features that are most likely to have been subject to pressure from conventionalisation and grammaticalisation processes in language. Instead, we want to examine features that capture differences between authors on a level of textual structure where the space for individual choice is wide: the organisation of informational flow and narrative frame. Such features can be obtained by studying configurations and progressions of observable properties above the clause level. We will call this family of aggregated features *configurational* in contrast to the typical pointwise measurements.

Rules, Constraints, and Conventions

The patent regularities of linguistic expression are formed by constraints – rules, conventions, and norms which can be of a biological, social, psychological, or communicative character. While language use is regular to a great extent, the rules that govern it change continously. 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 items are likely to have started their life as idiosyncratic choice, become accepted in some community as markers of some function, informational or social, and thence migrated to broader linguistic usage.

Given this progress 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 lexical or syntactic item, but between some *types* of observable items: text-global patterns, e.g. being less rule-bound than local clause-internal structure.

In brief, linguistic items grammaticalise, but first conventionalise. Some choices are optional, some non-optional. All such choices are not as accessible to the process of grammaticalisation. The linguistic items most studied in the fields of linguistics, information access, and stylostatistics 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.

The different levels of constraints are shown in Figure 1. 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.

SIGIR '07 Amsterdam. Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection

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 constructions. An observed divergence in a text sample from the expected occurrence of that specific item (with prior information 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 measurement only aggregates local statistics in texts and is 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.

Our hypothesis is that author (and speaker) choice on the level of informational structuring and organisation is less subject to pressure from conventionalisation and grammaticalisation processes. This both by virtue of wide scope, which limits the possibilities of observers to track usage, as well as the many degrees of freedom open for choice, which makes rule expression and rule following inconvenient.

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

advl the occurrence of more than one adverbial expression 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 the commonly used sentence length measure, and "advl" an estimate of topical elaboration and narrative anchoring of the text. An example analysis is given for reference in section .

Aggregation

Returning to the main claim of this paper, we investigate whether the introduction of configurational features spanning over text rather than local measurements might improve the potential for categorisation of authors. We wish to find an aggregation method which allows us to preserve some of the sequential information of author choice progression: as a candidate we measure the two features studied over a sequence of sentences, and record the resulting transition pattern for each feature over each text.

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: these transition patterns can then be compared for varying window lengths — the operational hypothesis being that a longer window length would better model variation over author rather than over genre.

Experimental data

We performed an experiment using

The method shown above example was performed on a larger set of genre-categorised texts. For the full 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 Figure 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.

ARTICLETYPE	n
advertising	522
book	585
correspondence	3659
feature	8867
leader	681
obituary	420
profile	854
review	1879
total	17467

Figure 2: Sub-genres of the Glasgow Herald.

Measurements and metrics

The measurements for the two chosen variables are given in Figure 3 for all genres and for some authors – the number of authors is large; only the authors with the highest and lowest scores for each variable are shown. The table shows, somewhat unsurprisingly, that the genres is more consistent with each other than are authors: some authors have really very few clauses ($clause_{min} = 0.52$) and very few adverbials ($advl_{min} = 0.39$) in their sentences, but all genres have a somewhat consistent density of subclauses and adverbials, spanning from 0.866 to 0.899 and from 0.601 to 0.735, respectively.

Transition patterns

To obtain the longitudinal patterns discussed above, each item, "clause" and "advl", was measured over sliding windows of one to five sentences along each text, and the occurrence of the feature was recorded as a *transition pattern* of binary occurrences, marking the feature's absence or presence in the sentences within the window. The first and last bits of text where the window length would have extended

feature	clause	advl
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 $clause_{max}$	0.96	
author $clause_{min}$	0.52	
author $advl_{max}$		0.81
author $advl_{min}$		0.39

Figure 3: Relative presence of features "clause" and "advl" in sentences.

over the text boundary were discarded. The feature space, the possible values of the feature with a certain window size is thus all the possible transition patterns for that window size. For windows of size two, the feature space consists of four possible patterns, for windows of size five, thirty-two, as shown in Figure 4.

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, yields 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 the category "correspondence", using a feature space for the feature "clause" 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, which can be seen from the last position in the vector below. Likewise, the first position of the vector tells us that the probability of finding three sentences in sequence with multiple clauses in a text in this category is 0.0069:

$$p_3(correspondence) =$$

$$= \{p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000}\} =$$

$$= \{0.0069, 0.0654, 0.00903, 0.155, 0.00454, 0.0363, 0.0486, 0.674\}$$

Evaluation

In categorisation tasks, an unknown item – in this case, a text – 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 assessement 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 the categories are clustered together. Or, in other words, one wishes to find features which separate categories well. So, given a particular 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. 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 working as a discriminating measure between the categories under consideration.

The cateory set is vastly different for authors and genres. As 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 horizontally in the table — the divergence figures for different window sizes (the rows of the table), cannot directly be related to each other, since the feature spaces are of different size. This means that we cannot directly say if window size improves

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".

Window size	Genre		Author	
	"clause"	"advl"	"clause"	"advl"
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 "advl".

the resulting representation or not, in spite of the larger divergence values for larger window size. Bearing that caveat in mind, 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: a larger window size allows a better model of individual choice than a shorter one.

Secondly, we find that the feature "advl" 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 better informed feature sets for the larger contextual models to improve distinction between individuals rather than genres.

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 lenghtier text spans might give better purchase for finding features open to 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 resulting distribution can be assumed to be up to individual choice rather than genre or topical convention or syntactic constraint.

The results of our experiment show that configurational features do give different results from pointwise features; they also support our contention that author categories and genre categories should be identified and discriminated in different ways – in the one case, identifying conventions, in the other, avoiding them.

At this juncture, the task is finding more (and more informative) features and factors of the less-conventionalised levels of the linguistic system, measuring them, evaluating them, and understanding and diagnosing their import on the knowledge representation we choose for an application. The features we expect to study are intended to reach beyond occurrence statistics, to measure presence or repetition rather than frequency, to avoid notions such as average and mean and instead to 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 competing knowledge representation schemes.

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 determine 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.

Example analysis

The following three texts describe the same event and were taken from various newsfeeds on August 26, 2007. Feature measurements are given in Table 7.

Example: Text 1

A powerful earthquake [jolted]$_{clause}$ eastern Indonesian islands [in North Maluku province]$_{advl}$ [Thursday]$_{advl}$, prompting government authorities to a tsunami warning. The quake, measuring 6.6 [on the Richter scale]$_{advl}$, [took place]$_{clause}$ [at about 0540 GMT]$_{advl}$, shaking Halmahera and nearby islands [in North Maluku province]$_{advl}$, [said]$_{clause}$ Fauzi, an official at Jakarta's Meteorology and Geophysics Agency. According [to the US Geological Survey USGS]$_{advl}$, the quake [was measured]$_{clause}$ [at 7.0 on the Richter scale]$_{advl}$. ”We have [issued]$_{clause}$ a warning that the quake [could [potentially]$_{advl}$ trigger a tsunami]$_{clause}$,” Fauzi [told]$_{clause}$ Deutsche Presse-Agentur dpa. He [said]$_{clause}$ the quake [took place]$_{clause}$ [about 57 kilometres beneath the seabed]$_{advl}$. No immediate casualties or injuries [were reported]$_{clause}$ [from the quake]$_{advl}$. Indonesia [is]$_{clause}$ located [in the Pacific volcanic belt]$_{advl}$ known as the ”Ring of Fire,” where earthquakes and volcanoes are common. [On December 26, 2004]$_{advl}$, a massive 9.0-magnitude earthquake, which [triggered]$_{clause}$ gigantic tidal waves, [devastated]$_{clause}$ thousands of homes and buildings [along the coastline of northern Sumatra]$_{advl}$, leaving around 170,000 people dead or missing [in Indonesia]$_{advl}$ and thousands more dead and injured [along the Indian Ocean coastline]$_{advl}$.

Example: Text 2

A powerful earthquake [rocked]$_{clause}$ eastern Indonesia [on Thursday]$_{advl}$, sending residents fleeing [from swaying homes and hospitals]$_{advl}$, authorities and witnesses [said]$_{clause}$. There [were]$_{clause}$ no immediate reports of damage. The quake, which [had]$_{clause}$ a preliminary magnitude of 7, [triggered]$_{clause}$ a tsunami warning but the alert [was]$_{clause}$ [quickly]$_{advl}$ lifted after it [became]$_{clause}$ clear no destructive waves [had been]$_{clause}$ generated, the country's geophysics agency [said]$_{clause}$. The earthquake [struck]$_{clause}$ [under the Maluku Sea]$_{advl}$ [at a depth of 20 miles]$_{advl}$, the U.S. Geological Survey [said]$_{clause}$ [on its Web site]$_{advl}$. The quake's epicenter [was]$_{clause}$ more than 130 miles [north of Ternate city]$_{advl}$. ”We [felt]$_{clause}$ a strong tremor [for almost a minute]$_{advl}$, people [ran]$_{clause}$ [in panic]$_{advl}$ [from buildings]$_{advl}$, [said]$_{clause}$ George Rajaloa, a resident in Ternate. ”Children [are]$_{clause}$ crying and their mothers [are]$_{clause}$ screaming, but there is no damage [in my area]$_{advl}$.” Indonesia, the world's largest archipelago, [is]$_{clause}$ prone [to seismic upheaval]$_{advl}$ [due to its location on the so-called Pacific ”Ring of Fire,”]$_{advl}$ an arc of volcanoes and fault lines encircling the Pacific Basin. [In December 2004]$_{advl}$, a massive earthquake [struck]$_{clause}$ [off Sumatra island]$_{advl}$ and triggered a tsunami that [killed]$_{clause}$ more than 230,000 people [in a dozen countries]$_{advl}$, including 160,000 people [in Indonesia's westernmost province of Aceh]$_{advl}$. [Just over a year ago]$_{advl}$, another quake-generated tsunami [killed]$_{clause}$ around 600 people [on Java island]$_{advl}$.

Example: Text 3

[According to the United States Geological Survey USGS]$_{advl}$ a strong magnitude 6.9 earthquake [has struck]$_{clause}$ Indonesia [in the Molucca Sea]$_{advl}$ [approximately 220 kilometers 135 miles north of Ternate, Maluku Islands, Indonesia]$_{advl}$ [at a depth of 44.6 kilometers 27.7 miles]$_{advl}$. The Japan Meteorological Agency [reports]$_{clause}$ the quake at a magnitude 7.0 with a depth of 50 km. An unnamed official with the USGS [says]$_{clause}$ ”there [is]$_{clause}$ a potential that a tsunami [might develop]$_{clause}$, [judging from the magnitude]$_{advl}$,” but no tsunamis [were]$_{clause}$ reported. ”We [have]$_{clause}$ lifted the warning. [After monitoring]$_{advl}$, there [were]$_{clause}$ no signs of tsunami,” [said]$_{clause}$ the Indonesian head of the country's geology agency, Fauzi.[Initially]$_{advl}$, Fauzi [issued]$_{clause}$ a tsunami warning saying ”we [have issued]$_{clause}$ a warning that the quake [could]$_{clause}$ [potentially]$_{advl}$ trigger a tsunami.”There [are]$_{clause}$ no reports of injuries, deaths or damage. One resident in Ternate [said]$_{clause}$ that he ”[felt]$_{clause}$ a strong tremor [for almost a minute]$_{advl}$, people [ran]$_{clause}$ [in panic]$_{advl}$ [from buildings]$_{advl}$. Children [are]$_{clause}$ crying and their mothers [are]$_{clause}$ screaming but there [is]$_{clause}$ no damage [in my area]$_{advl}$.” [Earlier]$_{advl}$ the National Oceanic and Atmospheric Administration NOAA [had issued]$_{clause}$ a tsunami bulletin stating that local high waves [could]$_{clause}$ be possible, but a widespread tsunami [is]$_{clause}$ ”not expected [based on historical earthquake data]$_{advl}$.”

	Text 1		*Text 2*		*Text 3*	
Sentences	8		10		10	
Words	175		213		203	
wps	6.6		6.2		6.2	
cpw	21.9		21.3		20.3	
clause	4		6		5	
advl	4		6		4	
1	-	+	+	+	-	+
2	+	+	-	-	-	-
3	-	+	+	-	+	-
4	+	-	+	+	-	-
5	+	-	-	-	+	-
6	-	-	+	+	+	+
7	-	-	+	-	-	-
8	+	+	-	+	+	+
9			+	+	+	-
10			-	+	+	+

Figure 7: Example texts, measurement of features

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 informmation 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.