Style in Text:
Creative Generation and Identification of Authorship

AISB '08

AISB 2008 Proceedings Volume 7

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008
University of Aberdeen

Volume 7:
Proceedings of the
AISB 2008 Symposium on Style in Text: Creative Generation and Identification of Authorship

Published by
**The Society for the Study of
Artificial Intelligence and
Simulation of Behaviour**

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 66 4

Contents

The AISB'08 Convention: Communication, Interaction and Social Intelligence

As the field of Artificial Intelligence matures, AI systems begin to take their place in human society as our helpers. Thus it becomes essential for AI systems to have sophisticated social abilities, to communicate and interact. Some systems support us in our activities, while others take on tasks on our behalf. For those systems directly supporting human activities, advances in human-computer interaction become crucial. The bottleneck in such systems is often not the ability to find and process information; the bottleneck is often the inability to have natural (human) communication between computer and user. Clearly such AI research can benefit greatly from interaction with other disciplines such as linguistics and psychology. For those systems to which we delegate tasks: they become our electronic counterparts, or agents, and they need to communicate with the delegates of other humans (or organisations) to complete their tasks. Thus research on the social abilities of agents becomes central, and to this end multi-agent systems have had to borrow concepts from human societies. This interdisciplinary work borrows results from areas such as sociology and legal systems. An exciting recent development is the use of AI techniques to support and shed new light on interactions in human social networks, thus supporting effective collaboration in human societies. The research then has come full circle: techniques which were inspired by human abilities, with the original aim of enhancing AI, are now being applied to enhance those human abilities themselves. All of this underscores the importance of communication, interaction and social intelligence in current Artificial Intelligence and Cognitive Science research.

In addition to providing a home for state-of-the-art research in specialist areas, the convention also aimed to provide a fertile ground for new collaborations to be forged between complementary areas. Furthermore the 2008 Convention encouraged contributions that were not directly related to the theme, notable examples being the symposia on "Swarm Intelligence" and "Computing and Philosophy".

The invited speakers were chosen to fit with the major themes being represented in the symposia, and also to give a cross-disciplinary flavour to the event; thus speakers with Cognitive Science interests were chosen, rather than those with purely Computer Science interests. Prof. Jon Oberlander represented the themes of affective language, and multimodal communication; Prof. Rosaria Conte represented the themes of social interaction in agent systems, including behaviour regulation and emergence; Prof. Justine Cassell represented the themes of multimodal communication and embodied agents; Prof. Luciano Floridi represented the philosophical themes, in particular the impact on society. In addition there were many renowned international speakers invited to the individual symposia and workshops. Finally the public lecture was chosen to fit the broad theme of the convention – addressing the challenges of developing AI systems that could take their place in human society (Prof. Aaron Sloman) and the possible implications for humanity (Prof. Luciano Floridi).

The organisers would like to thank the University of Aberdeen for supporting the event. Special thanks are also due to the volunteers from Aberdeen University who did substantial additional local organising: Graeme Ritchie, Judith Masthoff, Joey Lam, and the student volunteers. Our sincerest thanks also go out to the symposium chairs and committees, without whose hard work and careful cooperation there could have been no Convention. Finally, and by no means least, we would like to thank the authors of the contributed papers – we sincerely hope they get value from the event.

Frank Guerin & Wamberto Vasconcelos

The AISB'08 Symposium on Style in Text: Creative Generation and Identification of Authorship

This symposium aims to draw together researchers from two areas that have recently received attention: creative computing and forensic computational linguistics. We wish to foster interchange of experience and ideas from researchers interested in automated generation of text under various stylistic constraints, and researchers interested in identifying the authorship, origin or style of given texts. This can be viewed as two sides of a coin: creative generation or production of text; and identification of specific creative or distinctive factors.

Rodger Kibble
Sarah Rauchas

Symposium Chairs:

Rodger Kibble, Goldsmiths, University of London
Sarah Rauchas, Goldsmiths, University of London

Programme Committee:

Robert Dale, Macquarie University, Australia
Alexander Holt, Natural History Museum, London, UK
Rodger Kibble, Goldsmiths, University of London, UK
Richard Power, The Open University, UK
Sarah Rauchas, Goldsmiths, University of London, UK
Elizabeth Reeder, Glasgow University, UK
Graeme Ritchie, University of Aberdeen, UK
Maite Taboada, Simon Fraser University, Canada
Carl Vogel, Trinity College Dublin, Ireland
Geraint Wiggins, Goldsmiths, University of London, UK
Robert Zimmer, Goldsmiths, University of London, UK

Revisiting the Donation of Constantine

Francesca Frontini[1] and **Gerard Lynch**[2] and **Carl Vogel**[2]

Abstract. Techniques developed for synchronic text classification problems are applied to a significantly diachronic dataset. The scale of the temporal categories appears to matter. The problem addressed is that of using automated text classification methods to temporally locate *The Donation of Constantine*. The results reported do not contradict the analysis of Lorenzo Valla from 1440, claiming the document a forgery, but suggest that it is a very good forgery. This contributes to establishing the validity of these classification methods as applied to temporal categories and small datasets.

1 BACKGROUND

Some current work in computational linguistics returns to classical problems in historical linguistics [17, 7, 16, 18]. We apply text classification techniques using chronologically determined categories for a corpus of Latin texts. We report on experiments using letter bigram distributions as an index to Latin morphology and language change in Latin over time from extensive use of explicit case marking and free word order towards a period of scant variation in constituent order in written Latin. We focus on *The Donation of Constantine* (hereafter, the DOC) [2, 5].

By this name is understood, since the end of the Middle Ages, the alleged testament of Emperor Constantine the Great (272-373 AD), which is constructed as a bequest to the papacy, and used as a land claim by that institution. This document is without doubt a forgery, fabricated somewhere between the years 750 and 850. Its authenticity has been questioned all through the Middle Ages, but it wasn't until the XV century that its falsity was known and demonstrated. In 1433, in his *De Concordantia Catholica*, Nicholas of Cusa judged it as apocryphal. Some years later (1440) Lorenzo Valla (*De falso credita et ementita Constantini donatione declamatio*, Mainz, 1518 [5]) proved the forgery through the analysis of linguistic, stylistic and content anachronisms. Its authenticity was yet occasionally defended until Baronius, Cardinal and ecclesiastical historian, admitted in his "*Annales Ecclesiastici*" that the DOC was a forgery, whereafter it was soon universally admitted to be such [10, 12].

Although there appears to be a consensus that the text is a forgery, the document remains a focus of research into dating its language. For example, [6] is a recent work that examines the provenance of a phrase that appears in the document ("*urbis Romae episcopo et pape*"), whether this is best translated as "the bishop of the city of Rome and Pope" or "the bishop and pope of the city of Rome") due to the relative frequency of this description of the Pope in the 8th century and the resulting pragmatic force.

We revisit the question of dating the DOC using solely features of the text. The extent of text-external features is *a priori* classification of textual categories by the period of composition. The first analysis classifies the text using categories individuated by authorship (about 248 alternatives; plus 44 texts that are from anonymous sources), and the second considers the text with respect to seven temporal periods. The first analysis is interesting in revealing a 4th century historian as the source of text most similar to the questioned document. This does not contradict Valla, but suggests that as a forgery it potentially had a particular source from the target period as an exemplar. The second set of experiments divide the texts into broad temporal periods, extending to the contemporary period. This supplies additional support for claims of validity of the classification methods by suggesting that the text does not pattern well with Latin texts from 1400 to 1650, and less well with texts composed in the period from then until today.

The paper briefly outlines the methods which are tested here (§2) and describes the composition of the corpus (§3). In section §4, §5 and §6 we detail the experiments and demonstrate the outcomes. We conclude by indicating a range of other sorts of studies that we intend to explore specifically with respect to DOC, but more importantly in examining language change in Latin.

2 METHODS

For purposes of authorship attribution in forensic contexts, it has been suggested that letter unigram distribution analysis provides an anchor into the most reliable classification methods [3, 4]. The intuition behind reliability is the replicability of this level of tokenization: one may elect to count spaces, numbers, punctuation marks or not, but having made that decision there is no dispute about what counts as a letter. This is the opposite end of the spectrum from general linguistic "habits" that may figure into some methods of analysis [8]. Even hand tagging of part of speech gives rise to the need to assess inter-rater reliability. The more abstract the linguistic feature employed, the less agreement there will be about its application in individuating instances. This is in no small part because of gradience within linguistic categories [1].

One can conduct attribution studies with text-internal methods, external methods, or some hybrid [9]. External methods include reasoning about contemporaries and the descriptive content of the text, such as pointing out anachronisms. Here, we focus on the form of words in the texts, not the semantics.

The methods applied to assessing these distributions derive from suggestions about word-level tokenization from [14, 13] in connection with comparing corpora and, in that process, assessing corpus-internal homogeneity. The work here begins by collecting and individuating files of text. The input to the methods is an index of files and their *a priori* category. For the experiments reported in this paper, we balanced the file sizes at about 4KB of lines (using the unix split command). For these experiments, the natural category for a file

[1] Dipartimento di Linguistica, Università degli Studi di Pavia, Italy
[2] Computational Linguistics Group, Intelligent Systems Lab, Trinity College Dublin, Dublin 2, Ireland {vogel,gplynch}@tcd.ie

is the name of its author[3] or the period of Latin during which it was composed. We used letter bigram tokenization and record frequency distributions of letter bigrams in each file. We have used letter bigrams in order to capture the most productive unit of grammatical morphology encoding in Latin; however, for general classification letter unigrams have been proven quite useful [21, 11]. Pairwise similarity indices are constructed for all of the files using the cumulative chi-square value for each of the bigrams. The chi-square value is based on observed occurrences relativized to the file size (the normal computation of chi-square using observed values and expected values); this statistic is summed for each bigram that appears in either of the two files, and then divided by the number of degrees of freedom (essentially, the number of comparisons). In a divergence from the normal use of the chi-square test, we are not using the value to conclude that distributions are distinct or to reject a null hypothesis that they are random selections from the same population [15], but in order to rank their similarity. Thus, neither do we exclude from analyses comparisons with less observed frequencies less than the value five, as one ordinarily does when applying the chi-square as an adversarial test and not as a measure of similarity.

The rank ordering of similarity scores for each pair of files is then input to the Mann-Whitney rank ordering test to examine the significance of categories of files. For each *a priori* category, and each file within that category, we determine which categories the file fits best (the rank order of comparisons with the questioned file and the other files for each possible category is what one reasons with using the Mann-Whitney). Thus, each of the files within a category may or may not come up as significantly well suited to its *a priori* category and also with respect to other categories, simultaneously [20]. For the experiments reported here we are very generous in considering even $p < 0.25$ significance attached to similarity of a file with respect to a category as a relevant level of similarity. We use an additional level of significance testing for assessing category homogeneity—this is in thinking of a testing the fairness of c-sided coins, where c is the number of categories and the number of tosses is determined by the number of files in a category; thus, Bernoulli schema let us decide whether sufficiently many files are attributed to a category to deem the category homogeneous [19]. As sometimes the best fit category for one or more files of an *a priori* category is instead some other category, it is useful to also consider those alternative assignments.

3 CORPUS

The Latin Diachronic Corpus was sourced from the Internet. Latin sections are available on several text repositories on the web, such as Project Gutenberg[4] and Bibliotheca Latina at IntraText.[5] More specific sources are: `www.thelatinlibrary.com` and `www.documentacatholicaomnia.eu`. The primary source for the texts used here has actually been `www.thelatinlibrary.com` because it appears to be the most complete and with the widest diachronic span. Further, all of the texts are in the same html format. The documentation for the website indicates that

> "Many were originally scanned and formatted from texts in the Public Domain. Others have been downloaded from various sites on the Internet (many of which have long since disappeared). Most of the recent texts have been submitted by contributors around the world."[6]

The whole site has been downloaded using the program wget.

Pages not holding Latin texts were discarded. The html pages were automatically transformed into plain text:

- no accented or special characters
- no arabic numbers (to get rid of page numbers and other indexing numbers)
- no punctuation (which is spurious with respect to authorship in Latin, being added by the editor of the phylologic edition)
- substitution of U → V and u → v;

These changes are all intended to normalize the texts in a principled way. In modern Latin transcriptions, graphemes <u> and <v> were used to express the allophones of the unique old Latin phoneme /u/, which had an approximant pronunciation (such as in English <w>) and a plain vocalic one. Standards of transcription differ from text to text. This change prevents spurious heterogeneity. Making the substitution does not lose information, since the pronunciation is re constructable from phonemic context.

The files downloaded were renamed to indicate the age, authorship and work associated with each file. The temporal periods were:

1. archaic age:	early Latin text until 100 B.C.
2. classical age:	100 A.D. - 250/300 A.D
3. late imperial Latin:	300 - 600
4. early middle ages:	600 - 1000
5. high middle ages:	1000 - 1400
6. humanists:	1400-1650
7. modern and contemporary Latin:	1700 - today

The division between period 1 and 2 is quite clear from the point of view of language change. Whereas the first period is characterized by a higher degree of instability in the morphosyntactic features, as well as by archaisms in the lexicon, period 2 witnessed the creation and establishment of a standardized literary language whose rules were explicitly codified by grammarians and strictly followed.

Probably the most problematic division is the one between classical age (2) and late empire (3). From the literary and artistic point of view it is certainly objectionable to put classical authors such as Cicero, Virgil and Ovid (*latinitas aurea*), as we have done, in the same category with later authors (the so called *latinitas argentea*),[7] but from the point of view of linguistic change, especially of a highly conventionalized written language, it seemed plausible to create a broader category. The idea was to separate the mostly pagan classic and imperial Latin from the Christian literature which flourished in the late empire. The limit was set to 313 A.D., the year in which the Edict of Milan was issued by emperors Constantine and Licinius, proclaiming religious tolerance in the Empire, which set the path for Christianity to become a state religion.

Period 4 is characterized by the decay of Latin as spoken language, and by the emergence of Romance Languages. From the point of view of the literary language though, Latin established itself as the official language both of Church and Empire. Study of Latin revived especially in the IX century under Charlemagne (the so called "Carolingian Renaissance"). The boundary between 4 and 5 is not only

[3] The name of the author and particular work would also have been reasonable.

[4] `www.gutenburg.org` – Last Verified, January 14, 2007.

[5] `www.intratext.com/LATINA/` – Last Verified, January 14, 2007.

[6] See `http://www.thelatinlibrary.com/about.html` – Last verified, January 13, 2007.

[7] The *latinitas aurea* is latin for "the Golden Age of Latin" (75 B.C. to 14 A.D.); the "Silver Age of Latin" refers to the later period from from 14 A.D. to 200 A.D.

motivated by historical reasons—the new millennium being a conventional turning point from the early to the high Middle Ages—but also by linguistic ones. This period saw the coming of age of the new national languages used (both in the Romance as well as in the Germanic area) for the first time in the written medium as well. With the birth a literature in the modern languages, most of the authors belonging to this period became not only bilingual in speech (as already in period 4) but also in their written production (e.g. Dante).

As for period 6, it is known that Humanism is a movement, rather than a period, characterized by the rediscovery of the classics and the adherence to the Ciceronian style in the production; it spread from Italy and its influence reached the different parts of Europe in different moments between 1400 and 1650. In this category, we include authors who might show distinctive humanistic (that is classical) features in their style, thus avoiding the influence of the medieval category and upon the last category, which is composed by a series of very heterogeneous texts, dominated by the scientific Latin of the 18th, 19th century, as well as by the Latin of the church.

In this classification, the purported period for the Donation of Constantine is category 3 for late imperial Latin, and the accepted actual period of the text is category 4 of the early Middle Ages. However, finding similarity to the classical period is clearly relevant. No effort is made to balance the corpus by genre.

4 EXPERIMENT 1

Because the text of DOC amounts to only about 20K bytes, we individuated all of the files into approximately 4K chunks; this meant splitting DOC into six files. Removing files that were too small (less than 1K) left 19,963 files to be classified by author. In any run we examined five files of each category, 79 categories. Sources with too few files of sufficient size were eliminated in arriving at the 79 categories. This meant that we had balanced files by file size and categories for their number of constituent files. We analyze the aggregate results across twenty such random samplings. The homogeneity index for a category is the average number of times that files from the category fit best with the category itself.

Some of the more homogeneous categories are listed in Table 1. To consider significance for this one wonders how many tosses out of five coming up to the same side of a 79-sided die, in repeated experiments, it would take to conclude that the die is not fair: three is rather significant ($p < 0.005$). If the die is not fair, it is safe to reject the hypothesis that the files are clustering with their category because of random chance and accept that it is for reasons of meaningful similarity among the files. The rank ordering is based on similarity, after all. Note a preponderance of 4th century sources among the most homogeneous listed in Table 1. This might lead one to expect this temporal category to be the most heterogeneous. This is not the case, as is shown in the next section.

Thus, clearly the DOC is self-homogeneous as a document to have so high an index (2.737) after so many runs when being considered with respect to over 300 sources. However, recall from §2 that it is useful to examine the categories which are best fits for files of some category when their *a priori* category is not the best fit. In the case of the DOC, it finds Albert of Aix, Addison, Cassiodorus, Gregory of Tours and William of Tyre one time each in total across the 20 samplings. However, the remarkable thing is that the texts find Ammianus as a most similar alternative source 13 times. Note that Ammianus is particularly homogeneous in these experiments. This is striking given that the category is divided into 248 files, in contrast with the six files of the DOC. With regard to the dating of the document, while this does not refute accepted claims that the text was composed during the early middle ages, it is very interesting that the most reliable alternative source dates to the late imperial period, in which the text was originally claimed to have been written. This could perhaps point to Ammianus a source of language directly influencing the forgery.

Table 1. Most homogeneous author categories

Source	Approximate Period	Homogeneity
Addison	19th C AD	5.0
Albertanus	13th C AD	2.526
AlbertofAix	12th C AD	3.738
Ammianus	4th C AD	4.632
Apicius	4th C AD	4.316
Arnobius	4th C AD	2.842
Bultelius	16th C AD	4.368
Claudian	4th C AD	3.211
Commodianus	3rd C AD	4.632
DOC	8th C AD	2.737
Descartes	17th C AD	4.053
Gestafrancorum	12th C AD	4.053
Juvenal	late 1 early 2nd C AD	3.263
Kempis	15th C AD	3.789
Kepler	16-17th C AD	2.895
Walter	12th C AD	2.842

5 EXPERIMENT 2

We also wished to consider the text from the point of view of *a priori* temporal categories. We used the seven categories described in §3. Naturally, the category corresponding to the DOC is skewed in size in comparison to the other categories, but it is still useful to examine its texts in this way.

5.1 Temporal Categories with the *Donation* Isolated

As in the first experiment we took samples of five files from each category, however we ran 1000 such experiments. Table 2 shows the results. It is unsurprising that the DOC is the most homogeneous since it is a single text split into six constituent files from which five are chosen in any one experiment, while the other categories have many more from among which to choose five.

Table 2. Homogeneity of Temporal Categories

Period	Homogeneity
1archaic	1.569
2classic	2.247
3late	0.452
4earlyMA	1.252
5highMA	0.771
6humanism	1.018
7modern	0.779
donation	4.99

It is again very useful to consider the alternative assignments. First of all, Table 3 demonstrates that the best alternative category for the category defined by the DOC, out of 1000 runs, finds the classical period seven times (the period close to the claimed time of writing) and the high middle ages, twice. We emphasize that both are extremely small figures: in 1000 experiments involving five files each, only nine occasions was a DOC file not most similar to the DOC as a category.

Table 3. Alternative fits for the Donation of Constantine

Period	Fits out of 1000
1archaic	0
2classic	7
3late	0
4earlyma	0
5highma	2
6humanism	0
7modern	0

5.2 Temporal Categories Including the *Donation*

The homogeneity values for the temporal categories in Table 2 suggest a replication of that experiment with just seven categories: once with the files of the DOC recorded in category 4, the high Middle ages where consensus dates the document; and once with the files recorded with category 3. With repeated sampling, again 1000 experiments, and these texts recorded with a larger category, such that the DOC will not figure into the random selection of each and every experiment (as the construction from this section forces). Significance for this construction is as follows: eight categories and five files within a category are selected; for three to be assigned to its *a priori* category is only approaching significance ($p < 0.15$), but four is statistically significant ($p < 0.01$). Thus, only the Donation as a category is homogeneous. We do not expect any great changes to the values recorded for the proper temporal categories in Table 2 for either of these permutations.

Consider the case in which the files of the DOC are categorized with those of the late imperial period. Table 4 shows that the resulting homogeneity of the revised categories is not significantly changed (with seven categories and five choices, here, four out of five is significant $p < 0.05$). Only the classical period approaches significance. Both the late imperial period and the early Middle ages increase in homogeneity when the texts of the Donation are considered part of the late imperial period. The same is true when the texts are added to the category of the early Middle Ages (see Table 5). However, here the effect is much smaller. Further, both the late imperial period and the early Middle Ages form more homogeneous categories when the files of the donation are added to the late imperial period than when they are added to the category of the Middle Ages. This suggests greater compatibility with the earlier of the two periods.

Table 4. Temporal Homogeneity with the Donation in 3late

Period	Homogeneity
1archaic	1.657
2classic	2.218
3late	0.520
4earlyMA	1.349
5highMA	0.917
6humanism	1.184
7modern	0.814

Then it makes sense to consider the succession of tables 6-12. Initially we discuss the results in the first column of alternative fits (“Exp 2.1”); the other two columns correspond to the repeated 1000 experiments with the files of the DOC in 3late (“D3”), and then the 1000 experiments with those files in the category 4earlyMA (“D4”), instead. As with Table 3 the entries are on the same scale: an entry like 453 “Fits out of 1000” experiments means that with 5000 selections of files from the period given by the table, 453 were alternatively assigned to the category provided by the row as a better fit than the four other files in the *a priori* selection of five files for that experiment. The bottom row in each of the tables provides the total sum out of the experiments out of 1000 samples of five files per category in which a file in the category that the table records (stated in its caption) some find a category given by the earlier rows as a better alternative fit than the category in the caption.

Table 5. Temporal Homogeneity with the Donation in 4earlyMA

Period	Homogeneity
1archaic	1.681
2classic	2.254
3late	0.495
4earlyMA	1.282
5highMA	0.94
6humanism	1.05
7modern	0.811

First notice that with the DOC as a separate category, in fact, the classical period is the best alternative for each of the temporal categories. This speaks to the influence of that period on the language, and apart from the immediately preceding and following periods its best alternatives distribute more or less evenly, although with greatest overlap with humanism. This is consistent with Latin being a dead language by the time of the 6th century; from that period Latin was learned from written texts rather than ambient spoken language. During the humanism period, classical texts were consciously sought and rediscovered, influencing further the Latin in use during the humanism period. The pattern of similarity of periods of Latin to earlier and later periods is consistent with accepted philological thought. Table 6 shows that the archaic period finds as its best alternative the immediately following classical period in nearly $\frac{1}{10^{th}}$ of the tests.

The viability of the DOC as an alternative temporal category peaks with the High Middle Ages (Table 10). This analysis supplies no evidence supporting the claim that the DOC belongs to the late imperial period. Support for assignment to the late imperial period is roughly equal to that of the early middle ages.

Table 6. Alternative fits for the Archaic Period

	Fits out of 1000		
Period	Exp 2.1	D3	D4
2classic	483	453	468
3late	15	15	27
4earlyma	92	92	80
5highma	48	53	67
6humanism	199	211	175
7modern	90	104	107
Donation	4	n.a.	n.a.
Total Alternatives:	931	928	924

Inspection of the D3 and D4 columns reveals no change in the trends observed for the Exp 2.1 column of each of the tables. The differences that do exist are most interesting for the results reported in Table 7, Table 8, Table 9, and Table 11. Where differences of interest occur, it is because the location of the DOC texts within either cate-

Table 7. Alternative fits for the Classical Period

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	2	2	4
3late	48	44	62
4earlyma	199	223	181
5highma	104	101	116
6humanism	299	331	304
7modern	186	181	195
Donation	31	n.a.	n.a.
Total Alternatives:	869	882	862

Table 8. Alternative fits for the Late Imperial Period

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	0	0	0
2classic	474	483	505
4earlyma	145	148	141
5highma	117	113	130
6humanism	116	150	121
7modern	103	102	101
Donation	43	n.a.	n.a.
Total Alternatives:	998	996	998

Table 9. Alternative fits for the Early Middle Ages

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	1	2	2
2classic	487	483	476
3late	40	34	42
5highma	139	146	184
6humanism	133	158	140
7modern	115	129	114
Donation	46	n.a.	n.a.
Total Alternatives:	961	952	958

Table 10. Alternative fits for the High Middle Ages

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	0	0	0
2classic	437	413	450
3late	37	34	41
4earlyma	209	254	223
6humanism	141	164	142
7modern	95	107	112
Donation	56	n.a.	n.a.
Total Alternatives:	975	972	968

Table 11. Alternative fits for the Humanism Period

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	1	2	4
2classic	600	592	597
3late	20	30	25
4earlyma	112	134	107
5highma	75	77	99
7modern	148	135	141
Donation	22	n.a.	n.a.
Total Alternatives:	978	970	973

Table 12. Alternative fits for the Modern Period

	Fits out of 1000		
Period	Exp 2.1	D3	D4
1archaic	2	1	0
2classic	560	546	548
3late	24	32	40
4earlyma	123	138	103
5highma	75	77	96
6humanism	192	187	182
Donation	6	n.a.	n.a.
Total Alternatives:	982	981	969

gory 3 or 4 cause those categories, or the bellwether categories of the classical period or humanist periods, to change in overall homogeneity. This is represented in these tables by heterogeneity in increase or decrease in alternative assignments of the relevant categories.

Consider the overall force of these results. Take Experiment 2.1 (the first column in each of Table 6 through 12) as a baseline. In this paragraph, we summarize the trends reflected in the bottom line of each table. Of interest is the set of trends in homogeneity for each of the *a priori* categories when the text of the Donation is folded into either the late imperial period or the early Middle Ages. When the numbers in these tables increase, it is a sign that the category in the caption of the table has increased in heterogeneity. This is because these tables reflect the best alternative fits for the captioned category. Table 6 shows that when the DOC is added to the late imperial period (D3), the archaic period becomes more homogeneous, but more still when it is added to the early Middle Ages (D4). In the case of the classical period (Table 7), the category becomes less homogeneous when the DOC is in the late imperial period, but more so when it is in the early Middle Ages. The late imperial period (Table 8) becomes marginally more homogeneous when the DOC is within it (D3), and remains constant if the DOC is in the early Middle Ages (D4). The early Middle Ages (9) becomes more homogeneous with the DOC in it (D4), but is still more homogeneous with the DOC in the late imperial period (D3). The high Middle Ages (Table 10) is more homogeneous when the DOC is placed in either of the two periods, but slightly more so with it in the early Middle Ages. The humanism period (Table 11) is also more homogeneous with the DOC in either the late imperial period or the early Middle ages than in the baseline, but here, more so with it in the late imperial period (D3). Finally, the modern period (Table 12) is also more homogeneous with the DOC in either of the categories, but much more so with the DOC in the early Middle Ages (D4).

The tables for the late imperial period (Table 8) and the early Middle ages (Table 9) are the most critical. Both the late imperial period and early Middle Ages are more homogeneous when the DOC is considered as part of the late imperial period, the period in which the DOC claims itself to have been composed. Table 8 shows that the late imperial period finds the early Middle Ages as a best alternative marginally more when the DOC is in the late imperial period than when it is in the early Middle Ages. On the other hand, the early Middle Ages (Table 9) finds the late imperial period as a best alternative slightly less when the DOC is part of the late imperial period, and marginally more when the DOC is part of the early Middle Ages. Thus, there is equivocal support for locating the DOC in either of the two periods—slightly stronger similarity with the late imperial period emerges.

6 EXPERIMENT 3

As a control, we conducted a comparable analysis isolating the text of Apicius, who actually wrote in the 4th century, the period we are using as our 3rd category, Late Imperial Latin (3late). Recall from Experiment 4 that this was one of the relatively homogeneous sources, albeit based on a small number of files (15). The motive for choosing the source is partly that it actually is from exactly the Late Imperial Period in which the Donation was claimed to have been written.

6.1 Temporal Categories with Apicius Isolated

As in isolating the DOC we first consider the text from the point of view of *a priori* temporal categories. We used the seven categories described in §3. Again, we ran 1000 experiments, sampling five files from each of the temporal categories plus Apicius as a category, each time. The DOC was left categorized with the early Middle Ages as per scholarly consensus. Table 13 shows the results. It is unsurprising that Apicius is the most homogeneous since it consists of fifteen files from which five are chosen in any one experiment, while the other categories have many more from among which to select five. There is little change in the significance of homogeneity of the temporal categories with the texts of Apicius isolated.

Table 13. Homogeneity of Temporal Categories

Period	Homogeneity
1archaic	1.636
2classic	2.143
3late	0.464
4earlyMA	1.235
5highMA	0.852
6humanism	1.035
7modern	0.751
Apicius	4.815

Table 14. Alternative fits for Apicius

Period	Fits out of 1000
1archaic	1
2classic	58
3late	1
4earlyma	43
5highma	34
6humanism	23
7modern	24

Table 14 is to be compared with Table 3. Notice that when it is considered on its own as a category, it has considerably more fits with alternative categories than the DOC. Its most frequent alternative fit is the classical period, and strikingly, its best alternative fit is rarely its natural category (3late), just as for the DOC on both points.

6.2 Temporal Categories Including Apicius

The next tables consider the homogeneity of the seven temporal categories with the DOC where consensus locates it, but with the texts of Apicius varying between the late classical period and the early Middle Ages. The significance values are as in Experiment 2. In Table 15, homogeneity of 3late is lessened with Apicius in that category. Table 16 shows that the homogeneity of 3late increases when Apicius is classified as belonging to the early Middle Ages, and that period decreases.

Table 15. Temporal Homogeneity with the Apicius in 3late

Period	Homogeneity
1archaic	1.624
2classic	2.222
3late	0.473
4earlyMA	1.268
5highMA	0.962
6humanism	1.231
7modern	0.839

Table 16. Temporal Homogeneity with the Apicius in 4earlyMA

Period	Homogeneity
1archaic	1.749
2classic	2.273
3late	0.524
4earlyMA	1.217
5highMA	0.880
6humanism	1.070
7modern	0.852

Tables 17-23 show the for the other seven periods what the frequency of best alternative fits were. The first column represents the best alternatives when Apicius is considered in isolation as a temporal category (note that, unlike the DOC, it is never a best alternative). The second column displays the best alternative frequencies when the texts of Apicius are considered as part of the late Imperial Period, and the third column shows the results when Apicius is treated as part of the early Middle ages.

Table 17. Alternative fits for the Archaic Period

	Fits out of 1000		
Period	Exp 3.1	D3	D4
2classic	476	460	474
3late	28	14	31
4earlyma	90	83	94
5highma	55	54	39
6humanism	179	214	183
7modern	97	98	96
Apicius	0	n.a.	n.a.
Total Alternatives:	925	923	917

Just as Experiment 2 did for the DOC, consider the results of analyzing the temporal periods with the texts of Apicius isolated as

Table 18. Alternative fits for the Classical Period

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	1	4	4
3late	50	49	69
4earlyma	206	193	203
5highma	111	121	111
6humanism	326	309	293
7modern	181	178	189
Apicius	0	n.a.	n.a.
Total Alternatives:	875	854	869

Table 19. Alternative fits for the Late Imperial Period

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	0	0	1
2classic	504	469	529
4earlyma	149	133	137
5highma	116	141	106
6humanism	137	149	114
7modern	92	105	109
Apicius	0	n.a.	n.a.
Total Alternatives:	998	997	996

Table 20. Alternative fits for the Early Middle Ages

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	0	2	1
2classic	508	479	501
3late	26	28	36
5highma	137	174	148
6humanism	167	172	145
7modern	112	110	134
Apicius	0	n.a.	n.a.
Total Alternatives:	950	965	965

a baseline. The question, for each temporal period, is whether it is more or less homogeneous with the texts of Apicius considered as part of the late imperial period or as part of the early Middle Ages. Again, as before, we first consider the ramifications of bottom line f or each table. The archaic period (Table 17) is marginally more homogeneous with Apicius in the late imperial period and slightly more so with those texts in the early middle ages. The classical period (Table 18) shows greater homogeneity with Apicius in the late imperial period than in the early Middle Ages, but both are improvements over the baseline. The late imperial period (Table 19) is only marginally more homogeneous than the baseline, regardless of which of the two periods the texts are placed in. The early Middle Ages (Table 20) is significantly less homogeneous than the baseline whether the texts of Apicius are in the late imperial period or in the early middle ages. The high Middle Ages (Table 21) are much more homogeneous with Apicius in the late imperial period than in the early Middle Ages, and the same is true of the humanism period (Table 22). The modern period (Table 23) decreases in homogeneity from the baseline, regardless of which period Apicius is placed in.

Table 21. Alternative fits for the High Middle Ages

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	2	0	0
2classic	451	431	432
3late	34	35	52
4earlyma	232	228	247
6humanism	147	163	132
7modern	131	103	112
Apicius	0	n.a.	n.a.
Total Alternatives:	997	960	975

Table 22. Alternative fits for the Humanism Period

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	1	0	0
2classic	618	584	600
3late	29	27	26
4earlyma	118	102	109
5highma	86	103	78
7modern	122	140	158
Apicius	0	n.a.	n.a.
Total Alternatives:	974	956	971

Table 23. Alternative fits for the Modern Period

	Fits out of 1000		
Period	Exp 3.1	D3	D4
1archaic	4	1	0
2classic	553	541	572
3late	34	20	36
4earlyma	114	116	106
5highma	87	100	82
6humanism	183	202	180
Apicius	0	n.a.	n.a.
Total Alternatives:	975	980	976

As in §5, it is the difference for the late imperial period and the early Middle Ages that is of most relevance. Within Table 19, one can see that with Apicius in the late imperial period, the number of times that the early Middle ages as a best fit alternative for the late imperial period as a category decreases (column D3), and fits with the high Middle Ages increase. If the texts are in the early Middle Ages (column D4), then the category for the late imperial period still has fewer best fit alternatives to the early Middle Ages, and also fewer to the high Middle Ages, but more to the classical period. Best fit alternatives for the early Middle Ages (20) as a category also change from the baseline: when Apicius is part of the late imperial period (D3), alternatives for the early Middle Ages decrease with respect to the classical period from the baseline, stay about the same for the late imperial period, and increase for the high Middle Ages. With the Apicius texts in the early Middle Ages (D4), that period finds slightly fewer best alternative matches to the classical period, more alternative matches to the late imperial period and more to the high Middle Ages. These results show that the late classical period is a good fit for the texts of Apicius: the late imperial period finds the early middle ages as a best fit slightly more often when Apicius is within the early Middle Ages than when it is within the late imperial period, and the early Middle Ages finds the late imperial period as a best alternative more when Apicius is placed within the early Middle Ages than when it is in the late imperial period. Nonetheless, it is striking that Apicius, a larger category than the DOC, is never the best alternative for any of the other categories, in any of its 1000 samplings of five files.

7 CONCLUSIONS

First we try to synthesize the results of Experiment 2 and Experiment 3. Both experiments constructed a baseline, by considering the 7 temporal periods and a sub-corpus considered in isolation. In both experiments we add the sub-corpus to the late imperial period (the D3 column in the tables) and to the early Middle Ages (the D4 columns).

Hypothetically, if the sub-corpus is placed in its actual time period, one might expect, relative to the baseline, the category for its time period to have a reduced number of best fits with the competing category, and one expects the competing category to have a reduced number of best fits with the actual category. Moreover, if the sub-corpus is placed in the incorrect time period of the two competitors, then, relative to the baseline, one could expect the actual category to have an increased number of best fits with the incorrect period, and one expects the incorrect category to have a an increased number of best fits with the actual category.

Examining a control, if Apicius is part of the late imperial period (and it is), then when it is placed in the late imperial period, it should, relative to the baseline, have a reduced number of best fits for the early Middle Ages (and it does) and one expects the early Middle ages to have a reduced number of best fits with the late imperial period (which it does not). If Apicius is placed in the early Middle Ages (which is incorrect), then, relative to the baseline, there should be a greater number of best fits for the late imperial period with the early Middle ages (there are not), and an increased number of best fits for the early Middle Ages with the late Imperial period (which there are). Thus, the control does not actually fit expectations. This points to the inconclusiveness of the sort of stylometric methods we are using.

However, carrying the argument through, if the DOC is part of the late imperial period, then when it is placed in the late imperial period, it should, relative to the baseline, have a reduced number of best fits for the early Middle Ages (and it does not) and one expects the Early Middle ages to have a reduced number of best fits with the late imperial period (which it does not). If the DOC is placed in the early Middle Ages, then, relative to the baseline, there should be a greater number of best fits for the late imperial period with the early Middle ages (there are), and an increased number of best fits for the early Middle Ages with the late Imperial period (which there are).

Thus, both Apicius and the DOC have dissociations from expectations, if we have reasoned expectations correctly, but the dissociations are in different places. On one hand this speakes to the temporal influence of one period upon another, independently, and on the other hand it points out that the study to date is simply incomplete. It suggests that there is as much reason to doubt the temporal attribution of Apicius as there is the Donation of Constantine, or equally that there is no more reason to doubt the claim of the Donation than there is the provenance of Apicius. Apicius was chosen from within the late imperial period because of its size, it is certainly not a random sample—it merits additional investigation in its own right.

The name ”Apicius” actually refers to a collection of Roman cookery recipes, usually thought to have been compiled in the late 4th or early 5th century AD and written in a language that is in many ways closer to Vulgar than to Classical Latin. In the earliest printed editions it was given the overall title De re coquinaria (”On the Subject of Cooking”), and was attributed to an otherwise unknown ”Caelius Apicius”, an invention based on the fact that one of the two manuscripts is headed with the words ”API CAE”. Recall that Table 1 records this sub-corpus among the most homogeneous, and more so than *The Donation of Constantine*, its actual content make it rather distinctive and perhaps not the most ideal on that basis to use as a control in a study like this.

This paper reports on results to date in automatic analysis of corpora constructed around diachronic categories. These experiments on *The Donation of Constantine* show that internal analysis alone is compatible with the possibility that the document is not a forgery. External analysis derives from consensus in the literature that it is a forgery. Thus, it must be a very good one from the point of view of morphological similarities, whether or not they were intended as such. The letter bigram analysis conducted here was intended to discern patterns of Latin morphology, and as a sub-lexical treatment it explicitly abstracts over textual features that one might consciously control. Authors tend to make lexical decisions, not orthographic ones. Most texts are not lipograms. It is worth pursuing a possibility that Ammianus provided the source language input that shaped the forger's concept of fourth-century Latin. There is however, substantial reason to doubt such direct influence. Although Ammianus was Greek and his native language was Greek, he composed History in Latin, as the work was intended for Roman readers. The work consisted of 31 books and earned the author a considerable reputation in his day. It maintained at least some of its popularity until the 6th century, but then fell into neglect and is not mentioned during the Middle Ages. His work, given scholarship methods of the time, would not have been natural for an 8th century forger to stumble upon.

To further these experiments we intend to replicate them with letter trigram distributions. We also intend to record results for letter unigrams and word unigrams. It would be appropriate to normalize letters to upper-case. One obvious orthogonal advance would involve expanding the corpus with works without contention in their association to Constantine. This would give the problem a closer semblance to an authorship attribution task than to a temporal location task. A greater amount of text-external reasoning could perhaps be implemented by restricting genre within the sampling.

This is one thread of our ongoing work in text classification with temporal categories. The corpus itself is useful in providing a source of data with which to test our classification methods with respect to other established claims in the literature, towards settling the validity of the methods that we have been exploring. A goal for the research is to reliably quantify certainty about attributions of texts to categories on the basis of text-internal considerations. This is a necessary exercise for forensic purposes if evidence from linguistic analysis is to be acceptable to courts of justice.

ACKNOWLEDGEMENTS

Science Foundation Ireland (RFP 05/RF/CMS002) has supported this work. We are grateful to Jerom Janssen for useful advice and discussion. We also thank Rodger Kibble, Sarah Rauchas and the anonymous reviewers for encouragement and helpful feedback.

REFERENCES

[1] Bas Aarts, *Syntactic Gradience: The Nature of Grammatical Indeterminacy*, Oxford University Press, 2007.

[2] Anonymous, *Das Constitutum Constantini (Konstantinische Schenkung). Textausgabe*, Monumenta Germania Historia: Fontes iuris Germanici antiqui 10, Hannover, 1968.

[3] Carole Chaski, 'Linguistic authentication and reliability', in *Proceedings of National Conference on Science and the Law*, pp. 97–148, (1999).

[4] Carole Chaski, 'Empirical evaluations of language-based author identification techniques', *Forensic Linguistics*, **8**(1), 1–65, (2001).

[5] Christopher Coleman, *The Treatise Lorenzo Valla on the Donation of Constantine: Text and Translation*, New York: Russell & Russell, 1971. First published 1922.

[6] M. J. Edwards, 'Constantine's donation to the 'bishop and pope of the city of rome'', *Journal of Theological Studies*, **56**(1), 115–120, (2005).

[7] Mark Ellison and Simon Kirby, 'Measuring language divergence by intra-lexical comparison', in *Proceedings of the ACL*, pp. 273–280, (2006).

[8] Jill M. Farringdon, *Analysing for Authorship*, Cardiff: University of Wales Press, 1996. With contributions by Morton, A.Q., M.G. Farringdon and M.D. Baker.

[9] Don Foster, *Author Unknown. On the trail of Anonymous*, Macmillan: London, Basingstoke and Oxford, 2001.

[10] H. Fuhrmann, 'Konstantinische Schenkung', in *Lexikon des Mittelalters. 5: Hiera-Mittel bis Lukanien*, ed., R. H. Bautier, 1385–1387, Artemis, Mnchen and Zrich, (1991).

[11] Patrick G. T. Healey, Carl Vogel, and Arash Eshghi, 'Group dialects in an online community', in *DECALOG 2007, The 10th Workshop on the Semantics and Pragmatics of Dialogue*, eds., Ron Arnstein and Laure Vieu, pp. 141–147, (2007). Università di Trento (Italy), May 30 – June 1, 2007.

[12] C.G. Herbermann, E.A. Pace, C.B. Pallen, T.J. Shahan, and J.J. Wynne, *The Catholic Encyclopedia;: An International Work of Reference on the Constitution, Doctrine, Discipline, and History of the Catholic Church*, Appleton, 1907.

[13] A. Kilgarriff and R. Salkie, 'Corpus similarity and homogeneity via word frequency', in *Proceedings of Euralex 96*, (1996).

[14] Adam Kilgarriff, 'Comparing corpora', *International Journal of Corpus Linguistics*, **6**(1), 97–133, (2001).

[15] Adam Kilgarriff, 'Language is never, ever, ever random', *Corpus Linguistics and Linguistic Theory*, **1-2**, 263–275, (2005).

[16] Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tank, and Martin A. Nowak, 'Quantifying the evolutionary dynamics of language', *Nature*, **499**(11), 713–716, (2007).

[17] John Nerbonne, T. Mark Ellison, and Grzegorz Kondrak, 'Computing and historical phonology', in *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, 1–5, Prague, (2007).

[18] Mark Pagel, Quentin Atkinson, and Andrew Meade, 'Frequency of word-use predicts rates of lexical evolution throughout indo-european history', *Nature*, **499**(11), 717–720, (2007).

[19] Carl Vogel, 'Corpus homogeneity and bernoulli schema', in *Mining Massive Data Sets for Security*, pp. 93–94, (2007). NATO Advanced Study Institute.

[20] Carl Vogel, 'N-gram distributions in texts as proxy for textual fingerprints', in *The Fundamentals of Verbal and Non-Verbal Communication and the Biometrical Issue*, eds., Anna Esposito, Eric Keller, M. Marinaro, and Maja Bratanic, 189 – 194, Amsterdam: IOS Press, (2007).

[21] Carl Vogel and Sandrine Brisset, 'Hearing voices in the poetry of brendan kennelly', *Belgian Journal of English Language & Literature*, 1–16, (2007).

A Hybrid Statistical-Linguistic Model of Style Shifting in Literary Translation

Meng Ji

Abstract [1]**.** The present paper presents an original inter-disciplinary study of style-shifting in literary translation, which draws upon methodologies and techniques from corpus stylistics and computational stylometry, and relevant sociolinguistic theories of style variation. Such an innovative approach to the literary translator's idiosyncratic use of language sets out to address one of the most difficult issues in textual stylistics, i.e. the cognitive rationale behind style-shifting in literary translation. Keywords: Textual statistics; Corpus stylistics; Computational stylometry; Style shifting; Multi-variant analyses; Context-motivated theory in literary translation; contrastive linguistics

1. Outline of the current study

It is argued in the present study that style-shifting in literary translation is a very complex phenomenon which requires a quantification of source text contextual information potentially explanatory to such an important creative process in literary translation, which has been rarely discussed in depth in past studies. To distinguish the cognitive nature of stylistic variation as a conscious strategy devised on the part of the translator from being a simple reflection of the translator's writing habit, the context-motivated theory (CMT) is formulated and put to test with primary linguistic data retrieved from a parallel corpus containing Cervantes's Don Quijote in seventeenth-century Castilian and its two modern versions in Mandarin Chinese.

The hypothesis-testing process has been greatly facilitated by the application and experimentation of statistical techniques which are widely used in social and behaviour sciences, despite that their productivity in text-based corpus stylistics remains largely under-explored. The interesting results obtained in the present study suggest that in the study of stylistic variation, which is a crucial representation of the creative nature of literary translation, an integral approach to the subject matter which combines the discriminating strength of quantitative statistical analysis with the explanatory power of adjacent sociolinguistic theories holds the key to a deeper understanding of the cognitive nature of the textual phenomena under investigation.

2. Context-motivated theory of style-shifting in literary translation

Style-shifting has always been closed associated with the study of context-motivated or proactive speech variation in sociolinguistics (see Labov, 1972; Bell, 2001; Eckert & Rickford, 2001), primarily in qualitative terms; while its exploration from a computational stylistic point of view seems to have been less discussed. However, the present paper will show that the topic of style-shifting may actually be further developed and better explained through the processing of quantitative linguistic data from purposely built corpora. From the outset, it should be pointed out that the current study differs essentially from a pure quantification of linguistic information in search of underlying patterns in translational texts; instead, it focuses specifically on the exploration of contextual factors or parameters which tend to characterize certain textual phenomena that have been highlighted in a previous statistical study as a result of their abnormal or unpredicted occurrence in the corpus texts (Ji, 2008).

In this sense, it may be said that the current paper distinguishes itself from many stylometry studies (Henderson, 1978; Hoover, 2001), for in the place of being concerned with the devising and improvement of techniques that may increase the sharpness or accuracy of authorship attribution implements, it aims to delve into the cognitive dimension of stylistic variation, i.e. context-motivated /background-foregrounding or habitual /unconscious, in literary translation.

The design of the current project is largely based on a hypothetical proposition which attempts to explain stylistic variation brought about by literary translators in their work, drawing on their indirect or assumed estimation of the contextual situation as depicted in the source text. It offers an alternative explanation to style-shifting and addresses the issue from a perspective that is somehow different from previous theories, such as Labov's attention-to-speech model (1972) or Bell's audience design (1984, 1997 & 2001). Its theoretical speculation is based on observations of a specific type of style-shifting, where the textual phenomenon is seen as a pragmatic strategy developed on the part of the translator.

While past studies have seen stylistic variation in literary translations as (H1) invariably source-text derivative and (H2) such dependence is supposed to be largely fixed at a linguistic level, it is argued here that holding the first hypothesis as true, H2 may not necessarily follow. That is because, in dealing with linguistically challenging or somehow intangible linguistic issues in literary translation, such as archaism in a typologically and diachronically distant language, more often than not, the translator may resort to contextual cues in the original text, such as audience, setting, the relative social status between the parties involved in a speech event, etc.,

[1] Humanities, Imperial College London, SW7 2AZ
m.ji05@imperial.ac.uk

rather than linguistic cues, in an effort to overcome the linguistic challenges in the historical source texts1.

A key set of terms lying at the heart of the proposed context-motivated theory (CMT) is the concept pair of dominant versus latent, which is put forward to describe contextual features that have either stimulative or alleviative effects on the translator's decision of using a certain linguistic device. For instance, in studying the use of archaism, it is easy to fellow that taking the surrounding audience as a variable of the contextual analysis, the presence of a listening audience in the source text may well trigger off a stimulative effect on the translator's mind for him to use an archaic idiom, which in turn will enhance the chivalry image of the enchanted knight. On the contrary, the absence of audience, or a private setting of communication tends to have an alleviative effect on the conscious-minded translator, for it is commonsense that in a private setting, it is unlikely for one to take on an ostensibly archaic tone for showing-off purpose. Following this line of argument, all the source text contextual features highlighted in the current study have been tagged manually as dominant or latent as an initial encoding for later statistical analyses.

Lastly, it is envisaged that in the final analysis, the either context-motivated or habitual nature of the stylistic variation detected in the literary translation will be sustained by two different outcomes of the statistical modelling. That is, on the one hand, if the statistical modelling of contextual factors as assisted by the multi-variant analysis technique has helped establish a statistical model in each subdivision in which the contextual factors as constituents of the constructed statistical dimensions are largely dominant, then we may argue that the detected style shifting has been a conscious decision made by the translator to explore an idiosyncratic profile of his own. On the other hand, if the result of the statistical test shows that there is a proved instability in the contextual features characterizing the style shifting, or in other words, the contextual features seem to linger between dominant or latent values, it may therefore be said that the stylistic variance is more likely to be due to an unconscious use of archaism by the translator or simply a reflection of his or her habitual writing habit.

3. Hypothesis testing

The multi-variant technique used is categorical principle component analysis, also known as CATPCA. It aims to build a statistical model with limited dimensions, usually two, on a wide range of variables. CATPCA may help arrange a large amount of original variables in a way that is susceptible to human observation in finding interesting patterns in quantitative textual analysis. This is a vital process in the exploration of primary textual data where the reduction of numerous variables into a limited set of statistical dimensions holds the key to a deeper understanding of the nature of the original dataset; and hence provides important clues to research questions raised around the fundamental structure of textual data as measured in a two or three dimensional space. Despite the wide use of CATPCA in social and behaviour sciences (Stevens, 1992; Thomas, 2004), the applicability of such modelling technique in corpus stylistics remains to be tested, in this sense, the research methods developed in the current paper will undertake some initial investigations into the potential productivity of CATPCA in corpus stylistics.

Of course, such a context-feature-motivated approach may only survive in the environment of literary translation, especially in translating historical texts; literary translation should be seen as a creative process in its own right, which thus allowing some flexibility and pragmatism in handling with certain issues that are hard to pin down.

In a previous study, it has been proposed that to facilitate a comparative study of the source text and its two modern Chinese versions, the first part of the Spanish novel, which contains fifty-two chapters, has been divided into ten thematic divisions. The use of multiple regression test has revealed that while in most subdivisions, there is a corroborated similarity between the two Chinese translations in terms of the occurrence of archaisms in the protagonist's speech, in the ninth subdivision, perceivable discrepancies seem to emerge, which then brings about the issue of style-shifting in Liu's work.

In line with the CMT, a highly desirable model of contextual patterns which may be applied to explain a translator's deliberate use of language consists in an apparent consistency of the occurrence of situational parameters with dominant values as opposed to latent values. With regard to the use of archaism, it may be argued that archaism as an important form of rhetorical device tends to be used in a formal rather than informal setting of communication, where the degree of formality of communicative settings may be reasonably quantified by various situational parameters assigned with their corresponding values. For example, the quantification of interrelation-centric factors such as formality of relationship (FR) between addressor and addressees may be divided into latent values and dominant values within the context of the current study on archaism. To be specific, the dominant value of FR may refer to a formal relationship between the protagonist and his listeners, while the latent value of the same factor may be construed as a rather informal relationship between the communicating parties.

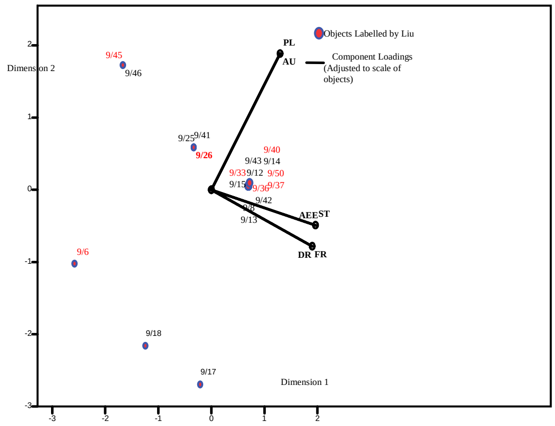
Table I Summary of VAF of Subdivision IX (Liu)

Total (Vector Coordinates)			
	Dimension		Mean
	1	2	
AU	0.319	**0.677**	0.996
ST	**0.732**	0.046	0.778
DR	**0.687**	0.116	0.802
FR	**0.687**	0.116	0.802
PL	0.319	**0.677**	0.996
AEE	**0.732**	0.046	0.778
Active Total	3.477	1.677	5.153
% of Variance	**57.948**	**27.943**	**85.891**

N.B. AU= Audience; ST= Social Status; DR=Distance of relationship; FR= Formality of relationship; PL= Place of communication; AEE= Addressee's evaluation of the content of speech;

Table I presents a summary of the two dimensional CATPCA model built upon the data retrieved from Liu's translation of Subdivision IX. As may be seen, the first dimension of the statistical model is primarily defined by the four interrelation-focused factors which are the distance of relationship (DR), formality of relationship (FR), relative social status of participants (ST) and finally, addressees' evaluation of the speech content (AEE). Meanwhile, the second dimension of the model is reduced to the two variables quantifying the communicative environment in which the protagonist has chosen to use archaic speeches or otherwise: audience and place of communication. As shown in the corresponding object-component-loading biplot of the statistical model, important patterns bearing on the distributional characteristics of IARS2 in relation to those of LTAS3 in Subdivision IX seem to emerge.

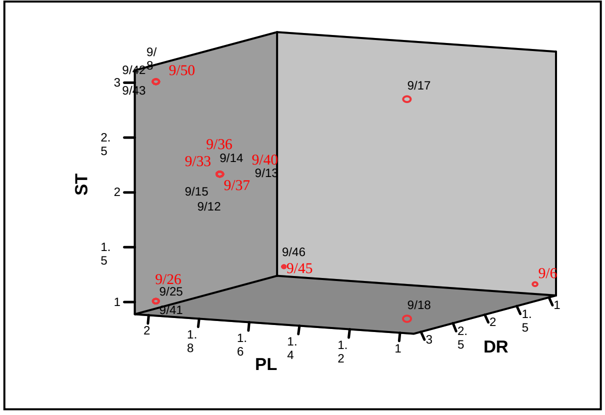
Diagram I Biplot of Subdivision IX (Liu, I)



Most object points in Subdivision IX, including IARS and LTAS, are clustered in the triangle space delineated by the two sets of variables functioning as the first and second dimension of the model. The clustering position is characterized by its restricted projection on the second dimension and its lower-intermediate value reported on the first dimension. To facilitate the discrimination of IARS from LTAS, the former has been marked in bold, which as the graph suggests, is nested intensively within the area demarcated by its LTAS counterpart and is thus quite difficult to separate them apart. There are some outliers running out of the concentration area and distributed sparsely in the biplot, which in the case of IARS are 9/6, 9/26 and 9/45.

There are some outliers running out of the concentration area and distributed sparsely in the biplot, which in the case of IARS are 9/6, 9/26 and 9/45. However, an important commonality shared by these IARS outliers is that they all feature to the left of the origin, while the vectors invariably extend into the areas to the right of the centroid. This graphical quality of IARS outliers implies that their values on the factors quantifying the first dimension of the model, i.e. DR/FR and ST/AEE, may well be equally as low as those IARS clustered to the central-right of the graph.

Diagram II 3-D plot of the distribution of Don Quijote's archaic speeches in Subdivision IX (Liu, I)



In view of the inherent limitation, of a two-dimensional graph, which may cause the visual overlapping of actually quite distinctively distributed factors, we proceed to convert the biplot into a three-dimensional graph. It should be noted that such procedure does not entail a structural modification of the already set-up statistical model; rather, it serves as a kind of assistant visualizing tool which will allow us a more intuitive access to the complex data structure under investigation. The visualization function used is readily provided in the SPSS 15.0 version known as chart-builder. As part of the software requirements, only one variable has to be specified in each of the three axes on the spatial representation of the data structure.

In Diagram II, we may see that the tri-partite pattern of the distribution of situational parameters already seem to emerge, and the high uniformity in the quantification of variables sustaining each "branch" of the pattern greatly facilitates the construction of the three-dimensional graph. That is, we need no more than to select arbitrally one variable from each parameter set and subsequently fill it in the mould furnished by the chart-builder of the software. The result of the construction of the three-dimensional plot is shown in Diagram II. An important feature of Diagram II is that it clearly sets apart the virtually overlapped two clusters of objects on the biplot, which has been made possible through the erection of the third dimension and the following segregation of the two minimally discerned variable sets on the biplot, i.e. FR/DR, PL/AU. In fact, as shown in the 3-D graph, the super-cluster which appears to the central-right on the biplot is actually composed by two quite distinct clusters as evidenced by their different locations along the ST and DR scales on the 3-D plot. In terms of the distribution of IARS, which constitutes our main concern in the establishment of contextual patterns that may explain Liu's stylistic use of archaisms in his translation of Don Quijote's speeches in Subdivision IX, we can see that most of the highlighted scatters, i.e. IARS, may be covered by the X-Y plane on the spatial representation of the data configuration. The only exception is 9/6 and 9/18, whose high value on the z-scale implicitly requires the specification of the variable PL.

In spite of the fact that IARS in Subdivision IX seems to cover both the lowest and highest ends of the X and Y scales on the 3-D chart, it is also obvious that IARS in Subdivision IX mainly appear in a mixed setting of communication as quantified by their medium value on both the ST and DR dimensions of the graph. The central cluster, which is composed of the following items 9/12, 9/13, 9/14, 9/15, 9/33, 9/36, 9/37, 9/40 is indicative of a contextual environment where the relative status and the interpersonal relationship among the parties to the exchange tends to be underspecified. The graphical analysis fits well into the general description of the textual plot, which with the exception of chapters when the protagonist is absent from the narrative scene, has shifted from the knight-squire private communication setting in the previous subdivision to the public domain as provided by the inn, which is also the title of the Subdivision IX.

4. Conclusion

Table II Summary of the CMT analysis

Statistical Dimensions	Value	Stylistic significance
ST (AEE)	Medium to high	Dominant
DR (FR)	Medium to high	Dominant
PL (AU)	High	Dominant

Table II presents the results of the CMT analysis of style-shifting in Liu's translation. In quantifying such narrative background with a view to establishing the patterns of contextual features that seem to characterize the use of IARS in Subdivision IX, we can see that the four interrelationship-focused factors come to the fore, i.e. DR/FR, ST/AEE. From here, it may be inferred that in his increased use of archaisms in don Quijote's utterance as detected in Subdivision IX, Liu seems to be rather sensitive to the interrelationship between the protagonist and his addressees; the contextual factor pair which focuses on the speech environment, i.e. PL and AU, also shows an unambiguous high value. As a result, it may be argued that according to the CMT, style-shifting, which has been detected in Liu's translation of Subdivision IX, is very likely to have been a conscious decision made on the part of the Chinese translator. The interesting finding uncovered through the formulation and testing of the context-motivated theory has thus provided us valuable and plausible explanations into the rationale or cognitive mechanism behind complex textual phenomena such as style-shifting.

Reference

[1] Bell, A. (2001) "Back in style: Reworking audience design", in P.Eckert and J. Rickford (eds.) Style and Sociolinguistic Variation, Cambridge University Press

[2] Eckert, P & Rickford, J (2001) Style and sociolinguistic variation, Cambridge University Press

[3] Labov, W (1972) Sociolinguistic Patterns, University of Pennsylvania Press

[4] Stevens, J (1992) Applied Multivariate Statistics for the Social Sciences (second edition), Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

[5] Thomas, B (2004) Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications, Washington DC: Ameican Psychological Association

[6] Ji, M (2008) "Quantifying style in two modern Chinese versions of Don Quijote", Meta: Journal des traducteurs, 53 (4), Quebec: Les Presses de l'Université de Montréal

[7] Yang, Jiang (1978) Tang Ji He De (Don Quijote), Beijing: People's Literature Publisher

[8] Liu, Jingsheng (1995) Tang Ji He De (Don Quijote), Gui Lin: Li River Publisher

[9] Henderson, Michael M.T. (1978) "Stylistic Variation and Underlying Structure", in Journal of Linguistics, vol. 14, no.2, 179-82

[10] Hoover, D (2001) "Statistical stylistics and author contribution: an empirical investigation", in Literary and Linguistic Computing, 16 (4), 421-44

[11] Cervantes, Miguel (1605) Don Quijote de La Mancha (part I), Francisco Rico (ed.), Instituto Cervantes online version http:// vc.cervantes.es/obref/quijote/

Style Variation in Cooking Recipes

Jing Lin[1] and **Chris Mellish**[2] and **Ehud Reiter**[3]

Abstract. Human text by specific authors has varied characteristics. Even when different authors write about identical topics within a fixed genre under a predefined domain, their text will still be different. For example, in food recipes there are style variations between authors even with describing the same cooking actions and foods. This paper summarises style variations in food recipes written by different people.

1 Introduction

Nowadays it is common to use corpora of human-authored texts to acquire knowledge in natural language generation (NLG), such as STOP [10] and SumTime [12], and so on. To make sure the corpora are big enough and broad enough, the corpora generally contain texts from several authors. When corpora are analysed, the differences between authors are normally ignored. Reiter and Sripada [9, 8] have suggested that "There are substantial variations between individual writers, which reduces the effectiveness of corpus-based learning". Our ultimate goal is to build an NLG system which translates recipes written by one author into the style of another author. This requires a good model of what aspects of recipes differ between authors; the corpus analysis presented here is a step towards such a model.

In the NLG area, some researchers have investigated producing text with varied styles. Hovy [4, 5] discussed rhetorical goals that are "the existence of a level of organization mediating between communicative goals and generator decisions". Furthermore, Stamatatos etc. [13] can produce outputs based on different user requirements, such as written style, tone, and so on. None of these researches focuses on generating texts with the style from an author. In literary comparison, *stylometry*, which is a subcategory of *attributing authorship*, researchers try to identify an author's style by building a set of measurable patterns[14]. The purpose of *stylometry* is to identify the author of texts, instead of describing the writing characteristics. Most of the defined *style markers* of stylometry are not enough to identify writing characteristics. For example, Burrows [1] only considers the frequencies of a group of the top (typically 30 or 50) frequent words. Moreover, some researchers used character-level n-grams recently, such as Kešelj [6], Peng [7], and Clement and Sharp [2], to consider the impact of morphological variations of words.

We introduce our corpora based on authors in Section 2 and give two example recipes by different authors in Section 3. Section 4 describes style variations in food recipes, which are categorised in at the sentence level and at the recipe level. The final section is our conclusion in Section 5, which relates this work to the problem of generating texts with styles of authors.

2 Our Corpora

To catch style variations between authors, we need to collect good quality corpora. Certain genres are less suited to analysing writing differences, such as daily news and fictions. Burrow [1] suggests that "in fiction: The language of its dialogue and that of its narrative usually differ from each other in some obvious and many less obvious ways". The authors of daily news often narrate matters with quotes from involved people and daily news is normally subjective from its authors, which make the text varied. Texts in food recipes, on the other hand, have features which are narrative, imperative, and objective. Although these features partly limit text variations, they also highlight styles between authors.

Based on the recipe domain, we collected a set of corpora based on authors from the Internet and recipe books. Table 1 presents information about our corpora. Our corpora include two recipe books, *Recipes for Health Eating* [3] and *Food for Health* [11]. There is an amateur recipe author: CM from the website (www.cooks.com). We also collected a few corpora of professional cooks from the BBC recipe website (www.bbc.co.uk/food/recipes).

Table 1. Author corpora information

	Corpus Information		
Author Corpus	Number of Recipes	Total Lines	Total Words
Recipes for Health Eating Dewhurst and Lockie [3]	76	961	9212
Food for Health Roe [11]	113	1347	11791
CM CM (www.cooks.com)	48	537	6432
Ainsley Harriott Harriott (www.bbc.co.uk)	87	891	14217
Antony Worral Thompson Thompson (www.bbc.co.uk)	1026	7790	99791
James Martin Martin (www.bbc.co.uk)	1103	8996	119582
Hugh Fearnley-Whittingstall Fearnley-Whittingstall (www.bbc.co.uk)	77	932	12187

3 Two Example Recipes

We list two recipes as examples from our corpora to provide an idea what our recipes look like. Both of the recipes are fish pie recipes, but

[1] University of Aberdeen, UK, email: j.lin@abdn.ac.uk
[2] University of Aberdeen, UK, email: c.mellish@abdn.ac.uk
[3] University of Aberdeen, UK, email: e.reiter@abdn.ac.uk

they show many differences between one and the other. For example, Step 1 of Recipe One is delayed to Step 7 of Recipe Two. Since each cook intends to create recipes with his speciality, it is unlikely to find two recipes describing exactly the same dish. In the following examples, we omit the ingredient section of the recipes as these do not typically contain syntactically valid English sentences.

Recipe One Fish Pie (Dewhurst and Lockie, [3])

1 Switch on oven to 220C, 425F or Gas Mark 7 and grease a 1/2 litre ovenproof serving dish.

2 Collect ingredients.

3 Wash and peel potatoes and cut into evenly-sized pieces. Place in boiling water and cook for approximately 20 minutes until soft.

4 Drain and mash potatoes when cooked. Add low fat spread and stir thoroughly.

5 Meanwhile, rinse and dry fish and poach gently in milk until fish flakes easily (approximately 10-15 minutes).

6 Drain and flake fish. Retain milk for sauce.

7 Measure warm milk and if necessary make up to 75 ml with water.

8 Mix mashed potatoes, flaked fish and milk.

9 Wash and finely chop parsley and add to fish and potato mixture with a good pinch of grated nutmeg.

10 Place mixture in greased ovenproof serving dish and using a fork, mark a pattern on top.

11 Bake for 10-15 minutes until brown on top.

12 Wash tomatoes, cut into wedges and arrange neatly on top. Place lemon slice and parsley in middle.

Recipe Two Creamy Fish Pie (Fearnley-Whittingstall, BBC online recipes)

1 Place the fillets of fish in a medium saucepan. Add the milk, onion, carrot, celery, bay leaf, a couple of stalks of parsley and the peppercorns.

2 Place the pan on a low heat and let the milk heat up gently. As soon as it comes to a simmer, switch off the heat and cover the pan. The fish will continue cooking in the hot milk.

3 Meanwhile, peel the potatoes, cut them into even, bite-sized chunks and put them in a large pan. Add just enough water to cover and put the pan on the hob over a high heat. Add a teaspoon of salt and let the water come to the boil. Lower the heat to a simmer and cook the potatoes until they are just tender.

4 Carefully drain the potatoes and allow them to cool in a colander for a minute or two. Return them to the pan and mash them, adding 50g/2oz of the butter, cut into cubes.

5 Stand a sieve over a large jug and tip in the fish and milk mixture. Wash the pan in which the fish was cooked and dry it well.

6 Add 3-4 tablespoons of the fishy milk to the mash and stir it in well. Add some freshly ground black pepper, taste the mash, and add some salt if you think it needs it. Put the mash to one side.

7 Heat the oven to 200C/400F/Gas 6.

8 To make the bechamel sauce, put the remaining 75g/2 1/2oz butter in the clean pan and melt it over a low to medium heat. Add the flour and stir well with a wooden spoon to make a roux. Cook for two minutes, stirring every few seconds. Then gently whisk in one third of the hot fishy milk. The paste will quickly turn into a very thick sauce. Add another third of the milk, whisking all the time, and then the final third, so you end up with a creamy sauce. Season the bechamel with salt and freshly ground black pepper, turn the heat down to very low, and let the sauce bubble gently for five minutes while you prepare the fish.

9 Remove the vegetables, herbs and peppercorns from the fish and discard. Carefully pick up a chunk of fish. Peel off any skin and discard, then gently feel the flesh between your fingers for bones, being careful not to over-shred the fish. Put the boneless fish on a clean plate.

10 Turn off the heat under the bechamel and add the fish to the sauce. Add the prawns, then chop the remaining parsley and stir this in too. Taste the sauce once more and add more seasoning, to taste.

11 Generously butter a pie dish and pour in the fishy bechamel. Spoon over the mash and spread it carefully across the surface of the fish sauce. Dot a little extra butter over the top of the pie.

12 Wearing oven gloves, put the pie in the oven and bake for about 25 minutes or until the top is starting to brown and the fishy sauce is bubbling up the sides of the mash.

13 Serve with buttered minted peas and crusty bread to mop up the sauce.

4 Style Variations

There are different types of style features with examples in food recipes. We catergorised them into sentence level and recipe level features. The style features at the sentence level are those that can be identified within one sentence, and do not have many connections outside of their sentence. The style features at the recipe level are those features which involve relationships between sentences, or cannot be simply defined within one sentence.

4.1 Sentence Level Features

Sentence level features include lexical preferences (Section 4.1.1), skipping objects (Section 4.1.2), content selection (Section 4.1.3), other differences at the sentence level (Section 4.1.4), and orthographical differences (spelling) (Section 4.1.5).

4.1.1 Lexical Preferences

Authors express the same action with different words, which can be synonyms with each other, or can be not. For example in Recipe One Step 1, Dewhurst and Lockie say *'Switch on oven'* whereas in Recipe Two Step 7, Fearnley-Whittingstall says *'Heat the oven'*. The following other examples show similar variation in the same action from other authors.

Example 1 "***Put*** the oven ***on***." (Roe, [11])

Example 2 "***Preheat*** the oven to 200C/400F/Gas6." (Thompson, BBC online recipes)

In the above four examples, these different verbs reflect the fact that the authors have different lexical preferences to express the same action — preheating the oven. Table 2 shows the frequency of the 4

Table 2. Sentence and word information of corpora of different authors

	Following Objects (oven or grill)			
Authors	Switch on	Put on	Preheat	Heat
Dewhurst and Lockie	35	2	0	0
Roe	0	42	0	0
Thompson	0	0	298	2
Fearnley-Whittingstall	0	0	4	7

verbs by the authors given objects *'oven'* or *'grill'*. Since *'switch on'* only appears in the authors, Dewhurst and Lockie, it can be considered as idiolect. In the recipe domain, idiolects may cover verbs, and nouns as well. For instance, some authors decide to use *'frypan'* instead of *'pan'*. Furthermore, Dewhurst and Lockie and Roe never use the word, *'preheat'*, in their recipes. Fearnley-Whittingstall's recipes are various, including desserts, main courses, and so on. Only small portion of his recipes involves bakeware, *'oven'* or *'grill'*.
A similar situation happens for other verbs, as in the following examples 3 and 4.

Example 3 "***Extract*** juice from orange and add this with the water to the saucepan." ([3])
Example 4 "Finely grate the ginger and ***squeeze*** out the juice into a shallow non-metalic dish." (Harriott, BBC online Recipes)

4.1.2 Skipping Objects

Some authors prefer to skip the objects of their verbs if they have been mentioned before, whereas others always include objects. See the following examples.

Example 5 "Blanch the sweet ***potato*** in boiling water for 2-3 minutes until softened. Transfer to a bowl and ***mash*** with the cream and herbs." (Thompson, BBC online recipes)
Example 6 "Wash and peel ***potatoes*** and cut into evenly-sized pieces. Place in boiling water and cook for approximately 20 minutes until soft. Drain and ***mash potatoes*** when cooked." (Step3 & Step4 of Recipe One) (Dewhurst and Lockie, [3])

Thompson always skips the object, *'potatoes'*, when he mentions the action, *'mash'*. Furthermore Fearnley-Whittingstall includes the object as a pronoun, *'them'*, in Step 4 in Recipe Two.

4.1.3 Content Selection

Authors present conditions using different information.

Example 7 "Bake until ***golden***." (Roe, [11])

In the above example, in Step 11 of Recipe One, and in Step 12 of Recipe Two, authors describe the appearance of the food when it is cooked, by using *'golden'* or *'brown'*. Both Dewhurst and Lockie and Fearnley-Whittingstall also describe the cooking time in their baking procedure. Fearnley-Whittingstall adds the delicious detail that *'the fishy sauce is bubbling up'* in contrast to the simpler descriptions by Roe and Dewhurst and Lockie. Roe uses subjective descriptions, such as *'golden'*, *'tender'*, and *'firm'* more often, instead of time information in her recipes. Her ways are more suitable for experienced cooks.

Some authors provide less details or no details, but others prefer to introduce actions with clear process descriptions. This happens especially in using certain verbs, like *'drain'* in the following examples.

Example 8 "Drain." (Roe, [11])
Example 9 "Drain." (Dewhurst and Lockie, [3])
Example 10 "Drain the beef from the marinade into a colander over a glass bowl." (Thompson, BBC online recipes)

When the authors present the action, *'drain'*, Roe and Dewhurst and Lockie rarely include further information, but Thompson most of the time presents detailed information related to the action, *'drain'*. This is in contrast to the example of *'mash'* in the previous section, in which Dewhurst and Lockie gave more information than Thompson.

4.1.4 Other Differences at the Sentence Level

Because of the freedom of the English language, authors can structure the same words into different sequences.

Example 11 "Peel and ***finely*** chop onion." (Dewhurst and Lockie, [3])
Example 12 "Peel and chop the onion ***finely***." (Roe, [11])

There are some unusual grammatical presentations in our corpora, for instance Dewhurst and Lockie always say *'Switch on oven'* in Step 1 of Recipe One, but never say *'Switch on the oven'*.

4.1.5 Orthographical Differences

There are some orthographical differences, which are recognised differently in computer systems. However, they should be considered as one word presenting in different forms, not two different words.

Example 13 "Pre-heat the oven to 220C/425F/Gas 7." (Thompson, BBC online recipes)

Thompson uses the word, *'preheat'*, in the form, *'pre-heat'*, 17 times. Moreover, Dewhurst and Lockie use *'Gas Mark 7'* in Step 1 of Recipe One instead Thompson just uses *'Gas 7'* in the previous example.

4.2 Recipe Level Features

At the recipe level, we find content differences (Section 4.2.1), order differences (Section 4.2.2), structure differences (Section 4.2.3), and aggregation (Section 4.2.4).

4.2.1 Content Differences

Some authors, such as Dewhurst and Lockie, always present certain cooking actions, but others, such as Thompson, skip these actions in recipe descriptions. Thompson never mentions *'collect ingredients'* and *'wash potatoes'* in the cooking method section, but Dewhurst and Lockie always does as showing in Step 2 of Recipe One.

On the other hand, most cooks only describe the food making process. However, Fearnley-Whittingstall also involves other actions. For example, he mentioned *'wearing oven gloves'* in Step 12 of Recipe Two and *'washing the pan'* in Step 5 of Recipe Two.

4.2.2 Order Differences

Different authors describe their recipes using a different order. In the Dewhurst and Lockie corpus ingredients are prepared when the time comes in the cooking process, so that the actions, *'wash and peel potatoes'*, are described in the cooking methods. Fearnley-Whittingstall follows this idea as well. On the other hand, Thompson assumes all ingredients should be prepared properly before cooking starts. For instance, *'1/2 sweet potato, peeled and diced'* in the ingredient section, instead of as Step 3 of Recipe One in the method part.

Describing preparing ingredients in the cooking methods is more suitable for experienced cooking learners, since they could handle many cooking processes at the same time. For beginners, Thompson's way could save a lot fuss.

4.2.3 Logic Preferences

In Section 3, we presented two recipes for fish pie, which contain common cooking actions. Table 3 shows the sequence of these actions appearing in the recipes. Since these two recipes are quite similar, it shows that the two authors have different logic preferences.

Table 3. Common cooking actions between two fish recipes

Action Code	Cooking Action
Action *a*	preheating the oven
Action *b*	making mash potatoes
Action *c*	cooking fish
Action *d*	baking the pie
Authors	**Action Sequence**
Dewhurst and Lockie	$a \rightarrow b \rightarrow c \rightarrow d$
Fearnley-Whittingstall	$c \rightarrow b \rightarrow a \rightarrow d$

4.2.4 Aggregation

Sometimes, one sentence in a cooking recipe contains two actions or more. This is called action aggregation. We expected to find different action aggregation habits from different authors. To our surprise, the action aggregations of different authors are unpredictable in our corpora. For instance, the authors, Dewhurst and Lockie, have aggregated the actions, *'switch on'* and *'grease'*, 10 times as in Step 1 of Recipe One, and have not aggregated them 20 times as in example 16. The author, Martin, also aggregated the actions, *'preheat'* and *'grease'*, 4 times as in example 17 out of 15 times. Since action aggregation is a common feature in food recipes, though it has no uniform style for each author, we feel it is worth mentioning here. Table 4 shows all the aggregation information about this section's examples.

Example 14 "***Switch on*** oven to 200C, 400F or Gas Mark 6 ***and collect*** ingredients." (Dewhurst and Lockie, [3])

Example 15 "***Switch on*** oven to 180C, 350F or Gas Mark 4. ***Collect*** ingredients." (Dewhurst and Lockie, [3])

Example 16 "***Switch on*** oven to 190C, 375F or Gas Mark 5. ***Grease*** a 1/ litre ovenproof serving dish." (Dewhurst and Lockie, [3])

Example 17 "***Preheat*** the oven to 170C/325F/Gas 3 ***and grease*** a muffin tray." (Martin, BBC online recipes)

Example 18 "***Preheat*** the oven to 200C/400F/Gas 6. ***Grease*** a baking dish with butter and dust with flour." (Martin, BBC online recipes)

5 Conclusion

In this paper, we collected and analysed many style variations between different authors. Some style features are at the sentence level. For example, most authors in our corpora have their own preferred words when they describe some actions in lexical preferences. Other style features locate at the recipe level, which rely on the subjective decision of an author. Logic preferences, for instance, clearly show the logic of an author.

In some style features, authors show clear preferences, as Dewhurst and Lockie always use *'switch on the oven'* and Martin uses *'preheat the oven'*. However, the aggregation between different authors are unpredictable. Another interesting observation is that authors include varied levels of details in their recipes, as Fearnley-Whittingstall even describes *'bubbling up'* in the baking step of Recipe Two.

Before a NLG system is capable to generate text with these writing style features, these style knowledge need to be extracted from corpus first in the knowledge acquisition process. Some style features require information from other style features. For example, all writing style features at the recipe level need information supports from lexical preferences and skipping objects. Therefore, to analyse the subtle style features, lexical preferences and skipping objects are the staring point in the knowledge acquisition. Only if all actions have been properly identified through lexical preferences, logic preferences, for example, can be identified.

Style features have their implications for NLG at different levels, if a NLG system intends to generate text with all these features. In the content determination, content selection at the recipe level and logic preferences need to be considered. Content selection at the sentence level and other differences involve in the sentence determination. Finally, lexical preferences, skip objects, and orthographical differences influence in lexical choice.

REFERENCES

[1] John F. Burrow, 'Word-patterns and story-shapes: the statistical analusis of narrative style', *Literary and Linguistic Computing*, **vol 2(2)**, pp 61–70, (1987).

[2] R. Clement and D. Sharp, 'Ngram and bayesian classification of documents for topic and authorship', *Literary and Linguistic Computing*, **vol 18(4)**, pp 423–447, (2003).

[3] Y. Dewhurst and G. Lockie, *Recipes for Healthy Eating*, Heinemann Educational Books, 1986.

[4] E. H. Hovy, *Generating Natural Language Under Pragmatic Constraints*, Lawrence Erlbaum, Hillsdale, NJ., 1988.

Table 4. Action aggregation Information about example 15 to example 19

Switch on — collect		Switch on — grease		Preheat — grease	
Aggregated	Separated	Aggregated	Separated	Aggregated	Separated
2	5	10	20	4	12
Total 7		Total 30		Total 16	

[5] E. H. Hovy, 'Pragmatics and natural language generation', *Artificial Intelligence*, **vol 43**, pp 153–197, (1990).

[6] V. Kešelj, F. C. Peng, N. Cercone, and C. Thomas, 'N-gram-based author profiles for authorship attribution', *Proceedings the Pacific Association for Computational Linguistics*, (2003).

[7] F. C. Peng, 'Language independent authorship attribution using character level language models', *Proceedings of the European Association for Computational Linguistics (EACL-03)*, (2003).

[8] E. Reiter and S. Sripada, 'Human variation and lexcial choice', *Proceedings of the 40th Meeting of the Association for Computational Linguistics (ACL-02)*, (2002).

[9] E. Reiter and S. Sripada, 'Should corpora texts be gold standards for nlg?', *Proceedings of the 2nd International Conference on Natural Language Generation (INLG-02)*, pp 97–104, (2002).

[10] Robertson R. Reiter, E. and Liesl Osman, 'Knowledge acquisition for natural language generation', *Proceedings of the First International Conference on Natural Language Generation (INLG-00)*, pp 207–215.

[11] D. Roe, *Food for Health*, Longman, 1990.

[12] Reiter E. Hunter J. Yu J. Sripada, S. and I. Davy.

[13] E. Stamatatos, S. Michos, N. Fakotakis, and G. Kokkinakis, 'A user-assisted business letter genertor dealing with text's stylistic variations', *the Ninth International Conference on Tools with Artificial Intelligence (TAI-97)*, (1997).

[14] N. Fakotakis Stamatatos, E. and G. Kokkinakis., 'Computer-based authorship attribution without lexical measures', *Computers and the Humanities*, (2001).

The Authorship of The American Declaration of Independence

Peter W.H. Smith, David A. Rickards

Abstract. Thomas Jefferson, the architect and author of the American Declaration of Independence (ADOI) is revered, and yet, even during his lifetime questions were raised about his authorship [1]. One name linked with the ADOI is Thomas Paine, the author of *Common Sense and The Rights of Man*. This study uses discriminant analysis to and Burrow's Delta scores [2] which reveal that both Jefferson and Paine exhibit a consistent style of writing. A word set is then created to discriminate between Jefferson and Paine using a hybrid genetic algorithm. From this an n-dimensional convex hull is used as test for authorship of the ADOI itself. Further tests are carried out based on sentence fragments. The study takes into account the sources of the ADOI including *The Virginia Constitution* and *The Summary View of The Rights of America* [8]. Results are based on an analysis of all known texts by Thomas Jefferson and Thomas Paine written *prior to* the signing of the ADOI. Test results indicate that Thomas Paine is the possible author of this historic document.

1 INTRODUCTION

Even during his lifetime, Thomas Jefferson was forced to defend his position as author of ADOI [1] .[3] is a very early attempt to use a form of systematic analysis to test the authorship of the ADOI.

At least seven different versions of the ADOI are known to exist [1], although fragments of an even earlier version have been found [4]. These versions are:

1. A Copy in the handwriting of Thomas Jefferson – known as *The Rough Draft.* (Massachusetts Historical Society, Boston).
2. A Copy in the handwriting of John Adams – thought to be an early copy of *The Rough Draft*. (Also at Massachusetts Historical Society, Boston).
3. A Further copy in the handwriting of Thomas Jefferson. (New York Public Library).
4. A Draft in the handwriting of Thomas Jefferson (American Philosophical Society, Philadelphia).
5. The Declaration as printed by Dunlap under the orders of Congress.
6. The Declaration written out in the corrected journal.
7. The Declaration on parchment at the Department of State.

At first glance, the ADOI looks unsuitable for an authorship attribution study, because such an important document would almost certainly have been much changed by Committee and The Continental Congress. Moreover, the evidence for Jefferson's authorship, would on the face of it, appear to be incontrovertible. He was appointed to draft the declaration and two texts known to have been written by Jefferson appear to be sources for substantial sections of the ADOI. However, a copy was produced prior to any amendments by Congress. Even the earliest known version appears to have been copied from an earlier copy now lost [1,4]. Furthermore, during his own lifetime Jefferson felt compelled to comment on his authorship stating that he drafted the ADOI, *without reference to pamphlet or book* - a comment that has puzzled scholars ever since.

The version chosen for this analysis is known as *The Adams' Copy*. There has been much speculation about it, though it seems clear, from a detailed analysis [1], that it is a copy of *The Rough Draft*. That it is in John Adams' handwriting, is not in dispute as his handwriting is confirmed in the biography by his grandson [5]. John Adams was also known to have sent this copy to his wife which also supports this view.

There are differences between *The Adams' Copy* and *The Rough Draft* that are puzzling. They are minor, as they consist largely of punctuation, but there are also a few minor spelling variations and couple of small grammatical changes. It has been suggested by [6] that the differences may be accounted for if *The Adams' copy* was written down by dictation. This simple explanation appears highly plausible, but in no way affects our study. Of the spelling variants between *The Adams' Copy* and *The Rough Draft*, none were found to be used by either Thomas Paine or Thomas Jefferson, but it was discovered that the spelling variant *tryal* was used by John Adams - in Clarendon No. 3 [7] - a small, but possibly significant finding that supports Whissell's theory.

The Adams' Copy was therefore chosen, because it is free of corrections made by Jefferson, Benjamin Franklin or by Congress and it is also possible, within reason, to date the evolution of the document. For the remainder of this paper, unless otherwise stated, we refer to *The Adams' Copy* of the ADOI.

2 METHOD

The study was carried out in four distinct phases:

- A study of the consistency of Jefferson's and Paine's writing using discriminant analysis and for comparison, Burrows' Delta method [2].
- The development of a word vector to separate the two authors, developed using a hybrid genetic algorithm.
- The application of the word vector to the ADOI.

[1] Department of Computing, City University, Northampton Square, London, EC1V 0HB. Email: peters@soi.city.ac.uk

[2] 18001 Euclid Avenue, Cleveland, OH 44112-1105, USA. Email: docrick@petalk.com

- A comparison using vocabulary, phrases and grammatical constructs.

2.1 The Writing of Thomas Jefferson and Thomas Paine

Thomas Jefferson's writing was extensive and varied. The authoritative source is [8]. For the first part of the study we chose a mixture of letters and documents written by him, in addition to his autobiography. Thomas Paine wrote a number of full-length texts as well as a series of articles. The authoritative source for his texts is [9]. For the first part of this study, we chose three of his full-length works as well as the set of essays known collectively as *The American Crisis*. This gave us seven texts for Jefferson and five by Paine, collectively over 500,000 words.

The Jefferson texts are as follows:

1. Jefferson letters and documents1760-1786 [j1]
2. Jefferson letters and documents 1786-1792 [j2]
3. Jefferson letters and documents 1792-1803 [j3]
4. Jefferson letters and documents 1803-1811 [j4]
5. Jefferson letters and documents 1812-1817 [j5]
6. Jefferson letters and documents 1817-1822 [j6]
7. Jefferson's autobiography 1821 [auto]

The following works by Paine were chosen:

1. Common Sense (1775) [sense]
2. The American Crisis (1776-1778) [crs1]
3. The American Crisis pt 2 (1778-1783) [crs2]
4. Rights of Man (1792) [rom]
5. The Age of Reason (1795) [aor]

The texts were converted into a standard form for consistency removing all extraneous text, quotes and foreign language sections. Jefferson's letters contain several passages in languages other than English and Paine's *Age of Reason*, for example contains a large number of biblical quotations.

For the purposes of the discriminant analysis, the texts were broken up into blocks of 2500 words. The discriminant analysis used the top 20 function words.

The results were checked by running Burrow's Delta method [2]. For this, we constructed a large text corpus of 2.4 million words, comprising texts by contemporaries of Paine and Jefferson, both British and American. The authors used in this corpus were: Alexander Hamilton, James Madison, Benjamin Franklin, John Adams, Adam Ferguson, Adam Smith, Edmund Burke, Edward Gibbon, Frances Brooke, Gilbert White, Horace Walpole, James Boswell, Robert Kerr, Samuel Adams, Samuel Johnson, Thomas Clarkson and William Beckford. The writing consisted of political speeches and documents, letters and longer texts, all of which are representative of the type of text written by Paine and Jefferson.

The discriminant analysis on the Jefferson texts revealed that texts j1/j2 and j4/5/6 were difficult to separate out. Additionally four of the five Paine texts showed a great deal of consistency of style by word frequency. However, one text for Jefferson and one for Paine appeared slightly problematic using this method – Jefferson's Autobiography and Paine's *Age of Reason.* The reason for the apparent difference of style in Jefferson's autobiography wasn't clear but Jefferson's writing spanned over 50 years and his autobiography was written in 1821. Paine's *Age of Reason* also exhibited different results, but it too was written later in Paine's life[3] - a considerable time after the ADOI.

To compute the delta scores, the method described in [2] was followed. The mean and standard deviation of the top 30 most frequent words was computed from the text corpus and the z-scores for all 12 texts were then computed. Then, taking each text in turn using it as an unknown against the other 11 texts, the delta scores were then computed. The results obtained using the Delta method were broadly similar. The results for Jefferson's early work, particularly pointed to a certain homogeneity of style.

2.2 Creating a Discrimination Word Vector

In the next phase of the study, we develop a word vector that is capable of accurately discriminating between Jefferson's and Paine's text. For this we used the 71 known texts by Jefferson written prior to the writing of the ADOI [8] and the 22 known texts written by Paine, also prior to the writing of the ADOI, [9]. We chose only texts written prior to the writing of the ADOI because Jefferson habitually quoted from the ADOI in his later writing. These texts were roughly equivalent in size forming approximately 54,000 words each. The texts were divided into blocks of 1000 words each. 32 blocks for each author were then used as a training set and the remainder used as a test set.

We rejected the use of Burrow's Delta method for two reasons: firstly it is primarily aimed at selecting the correct author from many and our study is aimed at selecting one of two authors, secondly, there are some methodological issues that are of concern. The Delta method relies on the Euclidean distance measure and it is well known that this is less effective over larger dimensional data sets. See for example [10,11] text mining applications that use large word vectors for document classification. We found a significant improvement in reliability with the Delta method after adapting it to use cosine similarity [12]. Although Hoover [13,14] reports improvements in the Delta method using larger word vectors, we were unable to replicate these improvements using the standard Euclidean distance measure, in fact our results suggest that performance of the Delta Method degrades with word vectors of size greater than about 200 words.

We felt that by simply choosing words by frequency particularly for very large word vectors was a coarse grained measure of style and better results could be achieved by selecting a word vector based on its ability to discriminate between two authors. The problem then becomes one of combinatorial optimisation as we need to select the optimal subset of words from the concordance listing.

We initially tried to evolve a discriminant function using a genetic algorithm, but then changed tactics following the completion of [15] in which subsets of words were chosen from the concordances of literary works by 4 authors to assess the degree to which different authors' texts exhibit clustering tendency. In a comparison of two texts by two different authors we measured the intra-cluster distance for the two clusters and compared this with the inter-cluster distance between the two clusters. These distance ratios were then compared using

[3] Age of Reason was also written during a difficult period in Paine's life – he was at that time condemned to death in France.

different word vector subsets of the concordance. This technique allowed us to investigate the clustering tendency of small word vectors which appeared to show promise of revealing why authorial style can be measured by word frequency. For example, it was noticed that some word pairs that clustered well also appeared together in for example, complex prepositions.

In addition to the intra and inter cluster distance measures we also computed an n-dimensional complex hull minimum enclosing space which acted as a means of testing unknown points in the n-space for authorship. We noted that for *n* blocks of data and an m-word vector, that *n>m*. This had the effect of dividing the space into three regions:

- The area within author 1's cluster.
- The area within author 2's cluster.
- The area outside both clusters.

This idea was developed in our current study. The choice of a subset of words to form a word vector from the words that appear in the joint concordance of Paine's and Jefferson's writing is a combinatorial optimisation problem. In order to find a good solution, we elected to use a genetic algorithm. The representation used was a simple binary string for which each bit position represented the presence or absence of a word in the chosen word vector. Bit 1 represented the most frequent word in the concordance, bit 2 the second most frequent etc.

For example in a concordance consisting of {the, and, to, of, from} – bit pattern 10101 represents word vector {the, to, from}. An initial population size of 5000 was chosen. Because of the constraint that the word vector should contain fewer words than data points, the initial population was seeded with vectors of different sizes ranging from 2 to 30 words. We elected to apply a large fitness penalty to any vectors containing more than 30 words. This resulted in a sparse bit pattern as the concordance contained more than 6000 words. Initially, the words that constituted the word vector in the initial population were chosen randomly. However, it was discovered that this approach created a population with a very poor overall fitness. It was long been established in authorship attribution studies that high frequency words are better indicators of fitness than lower frequency words. A typical concordance contains up to 60% of hapaxes, which are of dubious value as indicators of authorship. In a comprehensive study [25] also suggests that the most successful authorship attribution methods are based on high frequency words.

We then changed the method of initial word selection by biasing initial word choice according to the frequency with which the word appeared – a technique that biased selection to high frequency words, but did not rule out the inclusion of lower frequency words. This produced far better results.

Members of the population were then evaluated according to their fitness. Fitness being measured by the ability of the word vector to separate out the training set into two distinct cluster regions and to maximise the inter-cluster distance. We also discouraged the growth of larger word vectors by penalising word vectors over size 30 because of the constraint imposed due to the necessity to use word vectors of lower dimensionality than the number of available data points.

We initially used a standard method of crossover. However, because of the sparse nature of the word vector, we introduced a hybrid operator called "grow". This operator created an extra word in the word vector (i.e. it turned on a bit). Additionally, it worked significantly better by carrying out local hill-climbing on the grow operation in which the bits in the immediate area of the chosen bit (5 bits on either side) were tested for improvements to fitness. The resultant bit string was then added to the next generation.

Using this method, we found that optimal word vectors were created from vector sizes of between 25-30 words. As expected, these vectors were dominated by high frequency function words, but also contained a few words from the middle range of the Zipf distribution. Candidate word vectors were then tested on the test set for robustness.

Using this method, after several runs, we were able to create word vectors with completely disjoint clusters and with an accuracy of about 96% on the test set.

We were then ready for the next stage, which was to test the word vector on the ADOI itself.

2.3 Testing the Word Vector on the ADOI

In order to carry out an authorship study of ADOI, it is important to note the context in which it was written and also any sources that may have been used. Two texts in particular are important – *The Summary View of the Rights of America* written as *Draft of Instructions to the Virginia Delegates in Continental Congress* by Thomas Jefferson in 1774 and also *The Virginia Constitution*, thought to have been written in 1776 [8]. Nine textual similarities between SV and ADOI (i.e. passages that appear to have been directly taken or edited from SV) were identified and a further 16 similarities where passages in ADOI were clearly inspired by SV were also discovered. It was also noted that the 23 grievances against the King forming the central part of ADOI were either copied directly from, or edited from *The Virginia Constitution*[4]. Indeed 16 of these grievances appear in exactly the same order in both documents. It was therefore decided to analyse sections of the ADOI separately as it is clear that the central portion is either copied or edited from a text attributed to Jefferson. The ADOI was therefore subdivided into two parts:

- The List of Grievances (ADOIA)
- The Opening and Closing Statements of the Declaration (ADOIB).

We justify this on the grounds that the grievance list is copied or edited from *The Virginia Constitution*. ADOIB therefore contains 1104 words. ADOIA contained only 441 words. ADOIB was then tested using the word vector generated and tested over the Jefferson/Paine texts. The first word vector chosen was:

{ and, to, that, not, this, or, by, with, on, at, so, than, who, may, some, one, first, only, every, what, were, there, now, such, yet, same, when, out, had, up}

This word vector was then treated as a point in n-space and the position of this point in n-space was tested for inclusion within the Paine/Jefferson clusters defined using a convex hull algorithm for the minimum containing area of the cluster. The point was found to lie outside both the Paine and Jefferson

[4] There has also been some speculation as to whether the Virginia Constitution really preceded The ADOI [3]. Other documents have also been suggested as source for the ADOI, e.g. The Mecklenburg Declaration, which was commented on by John Adams, though it was later shown to post-date the ADOI [16].

cluster. The distance of the point from the computed centroids and the minimum distance from the Paine/Jefferson clusters was then computed. The minimum distance ratios for the Paine/Jefferson clusters was 1: 10.1, for the centroids it was 1: 8.8 – demonstrating that the point was far closer to the Paine cluster.

This exercise was repeated for 19 other word vectors all of which had high scores. In three cases out of 20, the vector for ADOIB was found to be inside the Paine cluster, in the other 17 cases, it was closer to the Paine cluster, using minimum distance, by a ratio of approximately 9.7 : 1. While this is not completely conclusive, we suggest that it casts doubt on Jefferson's authorship.

As a comparison, we applied Burrow's delta method to ADOIB against each of the files used in the first part of the study [2]. In Table 1, the lowest scores indicate authorship – in this case suggesting Paine as a likely author – but not conclusively. For comparison, we also applied the test to *The Summary View of The Rights of America* (table 2).

Table 1 Delta Scores for The Declaration of Independence

	j1	j2	j3	j4	j5	j6
ADOIB	1.529	1.564	1.69	1.721	1.709	1.637
	auto	aor	comm	crs1	crs2	rom
ADOIB	1.647	1.682	1.254	1.202	1.248	1.502

Table 2 Delta Scores for *Summary View of the Rights of America*

	j1	j2	j3	j4	j5	j6
Summary View	0.961	0.839	0.923	1.046	0.971	0.981
	auto	aor	comm	crs1	crs2	rom
Summary View	1.058	1.51	1.375	1.477	1.41	1.411

The results from table 2 indicate Jefferson as likely author of *Summary View*.

2.4 Further Tests based on Vocabulary and Grammar

In order to investigate the authorship of ADOI further, we carried out the following tests:

1. A Test of vocabulary usage based on words appearing in ADOIB.
2. A Test of Phrasal usage based on fragments taken from ADOIB.
3. A Test of Grammatical Usage based on the Function Words And/To

And/To appear in all word vectors that were capable of discriminating between Jefferson/Paine with high reliability. They were also high frequency words that differed markedly in their frequency in texts attributed to Paine or Jefferson. So these were investigated in much greater detail.

ADOIB consists of 441 different words, of which two are alternate spellings of the word "independent" in The Adams' version. Of these, 340 are hapaxes.

The study now concentrated on the set of texts by Jefferson/Paine written by them prior to the writing of ADOI. A search was conducted through all of these texts using all words that appeared in the ADOIB. The results, shown in Table 3 are not particularly conclusive either way.

Table 3 – Vocabulary Usage By Paine and Jefferson

	Jefferson	Paine	Both	Neither
Vocabulary	71	65	113	162
Hapaxes	38	41	98	143

For the next stage of the study, 344 phrases comprised of word collocations from ADOIB were constructed. The aim of this exercise was to determine the extent to which either author used phrasal fragments contained within ADOIB as well as the use of, for example, complex prepositions or prepositional verbs – something that might not show up in a word-based study. These fragments were then categorised as grammatical or content fragments. A search was conducted for the existence of the fragments in the works of Jefferson[5] and Paine written prior to the ADOI. Grammar-based approaches to authorship have been used elsewhere, for example [17,18] and particularly in forensic linguistics studies, for example [19,20]. [2] also differentiates certain function words by use of a part-of-speech tagger, albeit on a rather ad hoc basis

Initially, no distinction between the fragment types was made. The results are given (Table 4) in four categories as before:

1. Fragments unique to Jefferson.
2. Fragments unique to Paine.
3. Fragments used by both.
4. Fragments used by neither.

Table 4 – Fragment Usage Categories

Jefferson Only	Paine Only	Both	Neither
40	28	69	207

The fragments are now subdivided according to content or grammatical function: the results are given in Table 5.

Table 5 – Fragment Usage by Content/Function Word

	Grammatical	Content
Jefferson	16	24
Paine	16	12
Both	58	11
Neither	30	177

The results presented in Table 5 show that Jefferson scores proportionately higher on content but Paine scores proportionately higher on grammatical fragments. However, the

[5] The section of the Virginia Constitution dealing with Grievances was omitted because of its very close correlation with the central section of the ADOI.

distribution of the fragments unique to each author is also interesting.

ADOIA (the section of the ADOI consisting of the grievances) is considered next along with the equivalent section of the *Virginia Constitution* (Table 6).

Table 6 – Fragments Occurring in ADOIA

	VC	ADOIA
Jefferson	10	3
Paine	2	8
Both	8	6

13 fragments are unique to Jefferson and 10 are unique to Paine within ADOIA, these are proportionately about what would be expected, because ADOIA makes up about 30% of the total. However, a comparison of ADOIA and the *Virginia Constitution* reveals an interesting pattern: only two out of ten Paine fragments also appear in *Virginia Constitution*, whereas 10 out of the 13 of the Jefferson fragments also appear in *The Virginia Constitution*. The pattern indicates consistency of authorship for Jefferson for ADOIA and the VC. However, it also points to the influence of Paine in ADOIA, but not in the Virginia Constitution – possible evidence that ADOIA was edited by Paine based on the Virginia Constitution?

The software used in the search marked areas of the ADOI where matches occurred and the pattern of matches for Jefferson and Paine differed considerably. Jefferson's matches were clustered, whereas, Paine's were more or less evenly distributed throughout ADOIB and were less frequent in ADOIA. It was conjectured that this might be due to the fact that sections of the ADOI were edited from *Summary View*. The same search was then repeated having removed *Summary View* from the Jefferson text corpus. The results are given in Table 7.

Table 7 – Fragments Occurring in ADOI

	Grammatical	Content
Jefferson	16	24
Paine	16	12
Jefferson (minus Summary View)	8	14
Paine	16	12

The drop in both grammatical and content fragments clearly shows the influence of *Summary View* on ADOI – it also indicates that Jefferson's presence in ADOIB is in no small part due to *Summary View* being used as a source for ADOI. This in itself does not disprove Jefferson as author – but suggests an explanation as to why it is difficult to obtain a definite result on the authorship tests described above.

2.5 Jefferson and Paine's Use of "And" and "To"

Both Jefferson and Paine showed remarkable consistency with which they used high frequency words and it was noted that in particular, their usage of the third and fourth ranking words *and/to* was consistently different. It was also noted that the concordance of ADOI ranked the word *to* above *and* – a pattern consistent with Jefferson. However, if the central portion of ADOI (ADOIA) is removed, then this pattern reverses – though only just – making it more consistent with Paine. This called for a closer examination of the use of these words within the ADOI.

The most frequent words used by Paine and Jefferson are listed in Table 8. Both Jefferson and Paine consistently used *the* and *of* most frequently. However the frequency of their third and fourth words was consistently reversed. Jefferson used *to* more than *and* – the order is reversed for Paine. These words always appeared in the word vectors derived above. This ordering reversed for Jefferson later in life. In any authorship study it is important to use not only at methods that are capable of discriminating between authors, but also to attempt to understand why those differences exist. In this section we work towards a grammar-based analysis of Jefferson/Paine attempting to match this with grammatical constructs used within ADOI.

Table 8 – Jefferson/Paine Most Frequent Words

Word	Jefferson /1000 words		Paine /1000 words	
	Mean	SD	Mean	SD
the	59.94	8.74	69.00	13.56
of	44.04	8.12	46.38	5.70
to	40.08	7.78	30.76	6.67
and	26.73	5.97	35.81	4.44

Attention then focused on the use of these words, because of their potential to discriminate between Jefferson and Paine. It was also decided to examine why this difference between the authors existed. Function words *and* and *to* have multiple uses [21]. A recent study on forensic data [22] identified 16 different categories of use for the function word *to* and 38 different uses for the function word *and*. *To* is primarily used in either a *to-infinitive clause* or as a *preposition*. *And* is used mostly in *clausal co-ordination* and as a *sentence connector, a*lthough it also has other less frequently occurring uses.

The function word *and* is used in the following ways in ADOIB:

- As a connective with a following to-infinitive clause.
- To co-ordinate a binomial phrase.
- Use as a connective for phrases or sentences.
- Use as a complex connective.

To is used in the following ways in ADOIB:

- As a non-finite clause.
- In conjunction with an empty *it* clause with a non-finite clause.
- Beginning a sentence as a marker of an infinitive verb.
- As part of a complex connective *to which*.
- Various uses as a preposition, i.e. as part of a complex preposition, as a marker of a prepositional verb and as a simple preposition.

The uses of and/to were then used to construct a feature set to test whether the use of and/to matched Jefferson's or Paine's use of and/to.

Once again the texts written prior to the publication of ADOI were used. From the survey of usage of and/to in ADOIB, the following feature set was constructed:

- The use of *and* as a connective followed by a to-infinitive clause.
- *And* used in a co-ordinated binomial and phrase.

- The complex connective *and such.*
- *To* used as a non-finite clause.
- Constructions using the empty *it* subject with a non-finite clause.
- A *to infinitive* beginning a sentence.
- The use of the copula followed by a noun phrase.
- The complex connective *to which.*

And used as a simple connective and *to* used as a preposition were excluded from the feature set. Of the chosen features, it was discovered that two of them never occurred in either Jefferson or Paine's chosen writing. The remaining features were tabulated by help of custom written programs and a standard part-of-speech tagging program with manual checking to remove spurious matches. The results of this survey are given in Table 9. The row labelled J Total gives the total number of the feature found in Jefferson's writings. P Total is the total for Paine's writing. The frequency of co-ordinated binomial and phrases was so extraordinary that is worth a separate mention. Co-ordinated binomial phrases [23] pair words from all four major grammatical categories using *and/or* (for this study, only *and* is considered). *And* may co-ordinate noun and noun, e.g. *fish and chips*, verb and verb, e.g. *go and see*, adjective and adjective, e.g. *black and white* or even adverb and adverb, e.g. *slowly and deliberately*. It was noted that Thomas Paine used co-ordinated binomial and phrases with an unusually high frequency (10.9 instances per 1000 words). In modern English they occur with a frequency of 0.8 per 1000 words and a survey of 20 authors contemporary to Thomas Paine revealed that no other author used them with a frequency as high as Paine.

Five of the seven chosen features were distributed according to a poisson distribution. For this Mosteller and Wallace's classic study [24] was referred to. P-values and likelihood ratios were then computed (Table 9). The results of these also suggest Thomas Paine as the likely author.

Table 9 – Feature Set Chosen from the American Declaration of Independence

Feature	to-inf	empty-it + to inf	binomial and + to inf	and + to inf	binomial and
J Total	774	0	6	22	122
P Total	708	65	16	12	585
J Mean	15.6	0	0.06	0.41	2.3
P Mean	13.1	1.2	0.3	0.22	10.9
DOI Target	13	3	1	5	18

Feature	To-inf begins	to be +
J Total	9	27
P Total	15	44
J Mean	0.17	0.51
P Mean	0.28	0.82
DOI Target	1	1

Table 10 – P-values and Likelihood Ratios for the Poisson Distributions

Feature	Empty-it+ to inf	binomial and + to inf	and + to
P Value	P≥ 3	P ≥ 1	P ≥ 5
P val –J	0	0.0952	0.0001
P val –P	0.1203	0.2591	0
likelihood ratio J: P	1 : ∞	1 : 2.72	too small

Feature	To inf begins	to be +
P Value	P ≥ 1	P ≥ 1
P val –J	0.1813	0.3935
P val –P	0.2591	0.5507
likelihood ratio J: P	1 : 1.36	1 : 1.48

As the likelihood ratios show, once again the feature set based on the use of and/to suggests Paine as the likely author of ADOI.

Table 11 – Statistics For To-Infinitive Clauses and Binomial And Phrases

Feature	Statistic	Jefferson	Paine
to-infinitive	Total	774	708
	Mean	15.6	13.1
	DOI Target	13	13
	σ (standard deviation)	2.84	3.73
	No. of SDs from Target	<1	<1
	Range of Values	9-21	7-20
	Ranking Percentile	60-70	40-50
binomial and	Total	122	585
	Mean	2.3	10.9
	DOI Target	18	18
	σ (standard deviation)	1.58	4.64
	No. of SDs from Target	10	2
	Range of Values	0-6	5-24
	Ranking Percentile	>100	80-90

Table 11 presents the data for the two remaining features. The *to-infinitive* does not present strong evidence either way, but the *co-ordinated binomial and phrase* provides further evidence to support Paine, the mean and standard deviation for Jefferson is so low that it makes it highly implausible for him (more than 10 standard deviations away from the target) to be the author, on the other hand the data provides some additional evidence for Paine as the target figure of 18 for ADOIB is plausible as it is within the 80-90th percentile and is within two standard deviations of his mean score.

3 SUMMARY AND CONCLUSIONS

It has been demonstrated that Thomas Jefferson and Thomas Paine had consistent but different styles of writing using both Delta scores and discriminant analysis. It is argued that a method of authorship attribution for two authors should be based on an accurate word vector that is capable of discriminating between the two candidate authors with a proven level of accuracy. The choice of words for use is subset of the concordance – making it a combinatorial optimisation problem. A genetic algorithm with local hill-climbing was used to find a suitable word vector and found that word vectors between 20-30 words were perfectly adequate. This word vector was applied to a training set consisting of texts written by Jefferson/Paine prior to the writing of the ADOI. A test was also used to test reliability of the word vector. The test for authorship was based on using an n-dimensional convex hull algorithm, which created a minimum defining space for each author. Twenty different word vectors were then applied to the ADOI. In three cases, the ADOI was within the minimum defining region for Paine the remainder were far closer to the Paine cluster. Tests using phrase fragments, both content and grammatical also pointed at Paine as the potential author. Finally an examination of the grammatical use of *to* and *and* also suggest Paine as the more likely author.

The ADOI itself proved to be quite complex in that one substantial section was clearly copied and edited from another document and substantial sections of *Summary View* appear to have been edited into it. On the face of it, assuming the reliability of the tests used, there should be evidence of Jefferson's hand in its construction, but this is absent in our results.

However, it is clear that Jefferson's contribution to independence and the ADOI is considerable. It is suggested that Paine was asked or instructed to draft the framework of the ADOI by Jefferson, or by another member of the committee such as Franklin or Adams. It is almost certain that Paine was instructed to use material from both *Summary View* and *The Virginia Constitution*. While authorship attribution can never be entirely conclusive, the results presented provide a possible case for Paine's authorship.

Paine's commitment to independence is beyond doubt and he was in Philadelphia at the right time. His skills as a writer were well known to many members of the committee and he was known personally by Benjamin Franklin. He had already written anonymously and it would have been undesirable for an Englishman to have drafted the ADOI. There is also evidence that *The Rough Draft* was copied from an earlier version [1]. This adds plausibility that the ADOI was originally drafted by someone other than Thomas Jefferson, however, we have no doubt that The Adams' Copy was indeed drafted by John Adams.

It is also worth noting that some people have suggested that Thomas Paine must have been responsible for drafting the ADOI because of the inclusion of the infamous anti-slavery clause (omitted from the final version). This on its own is insufficient as references to anti-slavery are made in *Summary View*, though these appear to be more for economic than humanitarian reasons.

We would like to suggest that the findings of our research indicate that the possibility of Paine's hand in the drafting of the original version of ADOI and feel that this should be further explored.

ACKNOWLEDGEMENTS

We would like to thank Russell Gerrard for his comments on the statistical analysis and Sheila Munton for her interest in the project and for persuading the British Library to lend us several old texts. We would also like to thank Jane Riedel and George Collins for their comments and would also like to thank Andrew Tuson for his moral support for this research.

REFERENCES

[1] Hazelton, John H. The Declaration of Independence: It's History. Dodd, Mead & Co. (1906).

[2] Burrows J. 'Delta': A measure of Stylistic Difference and a Gudie to Likely Authorship. Literary and Linguistic Computing, 17,3,2002 pp. 267-287. (2002).

[3] Moody J. Thomas Paine: The Author of The Delcaration of Independence. John Gray & Co. (1872).

[4] Boyd Julian P. (1999) The Declaration of Independence: The Evolution of the Text. Ed. Gerard W. Gewalt. University Press of New England (1999).

[5] Adams, Charles Francis. The Life and Works of John Adams. Reprinted AMS Press (1856).

[6] Whissell C. http://www49thparallel.bham.ac.uk/back/issue9/whissell.htm. Accessed June 2005

[7] Thompson Bradley C. The Revolutionary Writings of John Adams. .Liberty Fund, Indianapolis (2000).

[8] Boyd Julian P. The Writings of Thomas Jefferson Vols. 1-27. Princeton University Press, Princeton, NJ. (1950).

[9] Conway, Moncure D The Writings of Thomas Paine Vols. 1-4 Reprinted by Ayer and Co.. (1894).

[10] Frigui H. and Nasraoui O. Simultaneous Clustering and Dynamuc Keyword Weighting for Text Documents. In Berry M.W. (ed.) Survey of text Mining: Clustering, Classification and Retrieval. Springer (2003).

[11] Korfhage R.R. Information Storage and Retrieval. Wiley, New York. (1977).

[12] Aldridge W.E. The Burrows Delta Dilemma: Optimisation of Delta for Authorship Attribution. M.Sc. Thesis. City University, London. (2007).

[13] Hoover D. Testing Burrows's Delta; Literary and Linguistic Computing 19,4, pp.453-475. Oxford University Press (2004).

[14] Hoover D. Delta prime?; Literary and Linguistic Computing 19,4,pp. 477-495, Oxford University Press. (2004).

[15] Smith P.W.H. The Clustering Tendency of Texts by Author. Unpublished. (2007).

[16] Hoyt, William H. The Mecklenburg Declaration of Independence: A Study of Evidence showing that it is Spurious. Da Capo Press. 1972.

[17] Baayen, R. H., van Halteren, H., & Tweedie, F. J. Outside The Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. Literary and Linguistic Computing, 2, 110-120, (1996).

[18] Lancashire, I. Phrasal repeatends in Literary Stylistics: Shakespear's Hamlet III.1. In S. Hockey & N. Ide (Eds.), Research in Humanities Computing. Selected papers from the ALLC/ACH Conference Christ Church, Oxford. Oxford: Clarendon Press. (1992).

[19] Grant, T. D. Reviewing and Revising Stylometric Authorship Attribution for use in a Forensic Context. Paper presented at the International Association of Forensic Linguistics 5th Biennial Conference, University of Malta. (2001).

[20] Smith Peter W.H. and De Jong G. Speaker Identification: Function Words and Beyond. Presented at The International Conference on Forensic Linguistics. Cardiff, July 2005. (2005).

[21] Quirk R., Greenbaum S., Leech G., Svartvik J. *A Comprehensive Grammar of the English Language*. Longman, London. (1989).

[22] Smith, Peter W.H. and De Jong G. Speaker Identification: Function Words and Beyond. Presented at The International Conference on Forensic Linguistics. Cardiff, July 2005. (2005).

[23] Biber D., Johansson S., Leech G., Conrad S and Finnegan E. *Longman Grammar of Spoken and Written English.* Longman, London. (2002)

[24] Mosteller F. and Wallace D.L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Prentice Hall. (1964)

[25] Grieve J. Quantitative Authorship Attribution: A History and an Evaluation of Techniques; Master of Arts Thesis, Department of Linguistics, Simon Fraser University (2005).

Three Approaches to Generating Texts in Different Styles

Ehud Reiter[1] and **Sandra Williams**[2]

Abstract. Natural Language Generation (NLG) systems generate texts in English and other human languages from non-linguistic input data. Usually there are a large number of possible texts that can communicate the input data, and NLG systems must choose one of these. We argue that style can be used by NLG systems to choose between possible texts, and explore how this can be done by (1) explicit stylistic parameters, (2) imitating a genre style, and (3) imitating an individual's style.

1 Introduction

Natural Language Generation (NLG) systems are computer systems that automatically generate texts in English and other human languages, usually from non-linguistic input data. For example, NLG systems can generate textual weather forecasts from numerical weather prediction data [8, 22]; descriptions of museum artefacts from knowledge bases and databases that describe these artefacts [15]; information for medical patients based on their medical records [5, 6]; explanations of mathematical proofs based on the output of a theorem prover [10]; and so forth.

NLG systems essentially have to perform three kinds of processing [19]:

- *Document Planning*: Decide what information to communicate in the generated text. This is usually based on an analysis of the information needs of the reader of the text.
- *Microplanning*: Decide how the chosen content should be expressed linguistically; that is, what words and syntactic structures should be used, how information should be packaged up into sentences, and so forth.
- *Realisation:* Create an actual text based on the above decisions which is linguistically correct, and in particular conforms to the grammar of the target language.

In this paper, we focus on the second choice, deciding how to express information. In most cases there are dozens (if not thousands or even millions) of ways in which a piece of information can be expressed. Making such choices is one of the least understood aspects of NLG, and we believe that models of style (interpreted broadly) can be very useful tools in making such choices.

2 SkillSum

In order to make the following discussion concrete, we will use examples from SKILLSUM [25, 31], an NLG system which was developed by Aberdeen University and Cambridge Training and Development Ltd. SKILLSUM generates feedback reports for people who have just taken an on-line screening assessment of their basic literacy and numeracy skills. The input to the system is the responses to the questions on the assessment (an example assessment question is shown in Figure 1), plus some limited background information about the user (self-assessment of skills, how often he/she reads and writes, etc). The output is a short report (see example in Figure 2), which is intended to increase the user's knowledge of any problems that he or she has, and (if appropriate) encourage the user to enrol in a course to improve his or her basic skills.

SKILLSUM must perform the three tasks described above. Briefly (see architectural description in Figure 3):

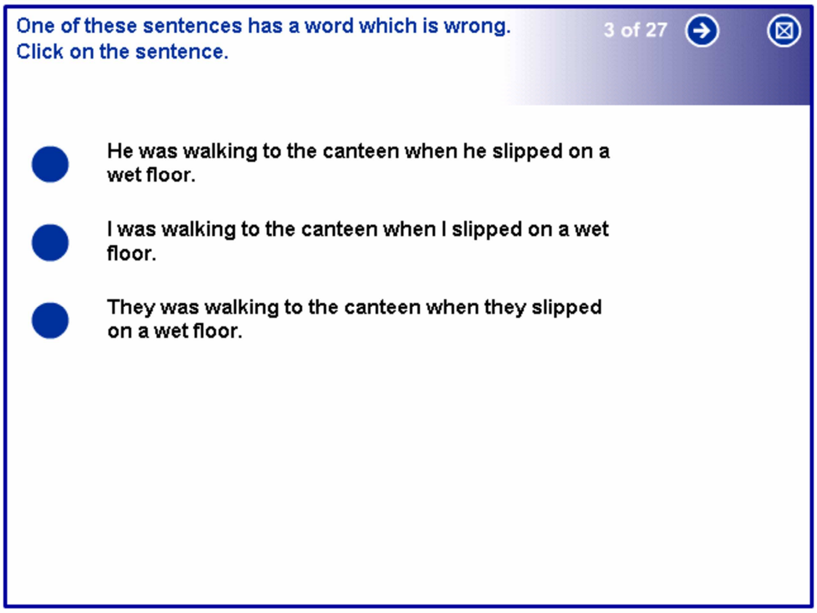
- *Document planning:* SKILLSUM uses schemas [13] to choose content. That is, it chooses content based on a set of rules which were originally devised by analysing and 'reverse engineering' a set of human-written feedback reports, and which were then revised based on feedback from domain experts (basic skills tutors) and also from a series of pilot experiments with users [29].
- *Microplanning:* SKILLSUM uses a constraint-based approach to make expression choices. The SKILLSUM microplanner has a set of hard constraints and a preference function [30]. The hard constraints specify which choices and which combinations of choices are linguistically allowed. The preference function rates the choice sets; SKILLSUM chooses the highest scoring choice set allowed by the hard constraints. As discussed below, style seems especially useful in the context of the SKILLSUM preference function.
- *Realisation:* SKILLSUM includes two realisers, one of which operates on deep syntactic structures [11], and the other of which operates on template-like structures

To take a simple example of microplanning, suppose that SKILLSUM wants to tell a user that he got 20 questions right on the assessment, and that this is a good performance. A few of the many ways of saying this are:

- *You scored 20, which is very good.*
- *You scored 20. This is very good.*
- *You got 20 answers right! Excellent!*
- *Excellent, you got 20 answers right!*

[1] Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK. Email: e.reiter@abdn.ac.uk

[2] Department of Computing Science, The Open University, Milton Keynes MK7 6AA, UK. Email: s.h.williams@open.ac.uk



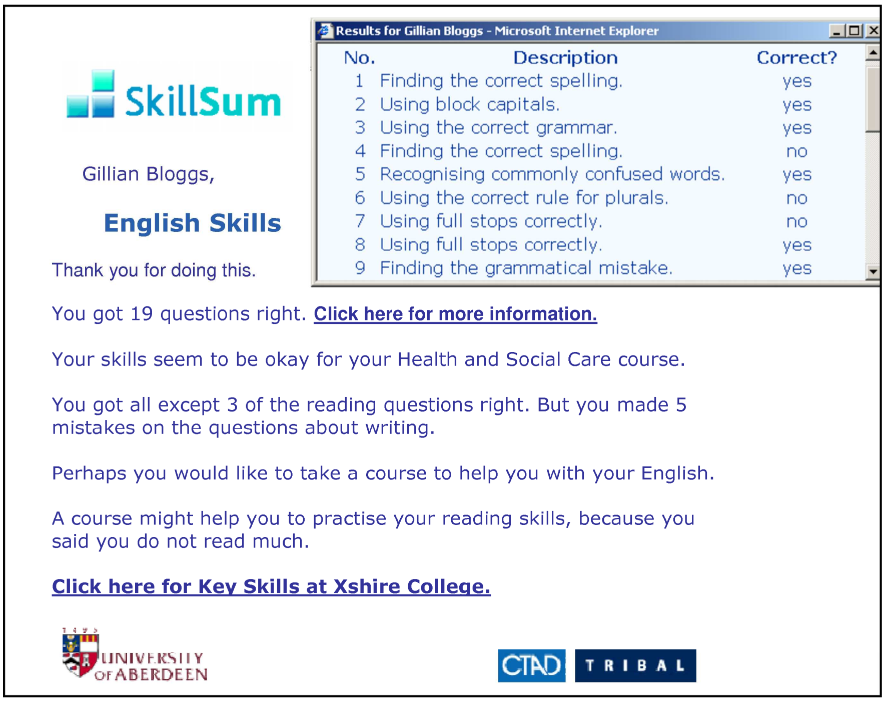
One of these sentences has a word which is wrong.
Click on the sentence.

● He was walking to the canteen when he slipped on a wet floor.

● I was walking to the canteen when I slipped on a wet floor.

● They was walking to the canteen when they slipped on a wet floor.

Figure 1. Example SkillSum Assessment Question



No.	Description	Correct?
1	Finding the correct spelling.	yes
2	Using block capitals.	yes
3	Using the correct grammar.	yes
4	Finding the correct spelling.	no
5	Recognising commonly confused words.	yes
6	Using the correct rule for plurals.	no
7	Using full stops correctly.	no
8	Using full stops correctly.	yes
9	Finding the grammatical mistake.	yes

Gillian Bloggs,

English Skills

Thank you for doing this.

You got 19 questions right. Click here for more information.

Your skills seem to be okay for your Health and Social Care course.

You got all except 3 of the reading questions right. But you made 5 mistakes on the questions about writing.

Perhaps you would like to take a course to help you with your English.

A course might help you to practise your reading skills, because you said you do not read much.

Click here for Key Skills at Xshire College.

Figure 2. Example SkillSum Output Text

- *20 questions were answered correctly, this is a very good score.*

The above examples illustrate some of the choices that are made in the microplanning process:

- *Lexical choice:* Which words should be used to communicate information? For example, should the first verb be *scored*, *got*, or *answered*?
- *Aggregation:* How should information be distributed among sentences? For example, should the above information be

communicated in one sentence or in two sentences?

- *Ordering:* What order should information be communicated in? In the above example, should the numerical score (*20*) or the qualitative assessment (e.g., *excellent*) come first?
- *Syntactic choice:* Which syntactic structures should be used? For example, should sentences be active voice (e.g., *You answered 20 questions ...*) or passive voice (e.g., *20 questions were answered ...*).
- *Punctuation:* For example, should full stops (".") or exclamation points ("!") be used?

The above list is of course not exhaustive; for example it does not include deciding on referring expressions (e.g., *The big dog* vs. *Fido* vs. *it*), which is not very important in SkillSum, but is important in many other NLG applications. Decisions also of course have to be made in the other NLG stages (document planning and realisation), but we will focus on microplanning in this paper. We will also focus on how style affects words, syntax, and sentences, and ignore how style affects visual aspects of text such as layout [17].

3 Using Style to Make Microplanning Choices

One appealing way to make decisions about lexical choice, aggregation, and so forth is to appeal to psycholinguistic knowledge about the impact of texts on readers. For example, if an NLG system is trying to generate texts which are very easy to read (as was the case with SkillSum), it would be nice to base choices on psycholinguistic models of the impact of different words, sentence lengths, and so forth on reading speed and comprehension [9]. Similarly, if an NLG system is trying to generate texts which motivate or persuade people (such as STOP [20], which generated personalised smoking-cessation letters), it seems logical to base these choices on psycholinguistic models of how texts motivate and persuade people.

Unfortunately, our knowledge of psycholinguistics is imperfect, which makes this difficult to do. Also in practice context (such as how much sleep the reader had the previous night) can effect the psycholinguistic impact of different choices; and such contextual knowledge is usually not available to NLG systems. SkillSum in fact tried to base some of its choices on psycholinguistic models of readability, and while this worked to some degree, overall this strategy was less effective than we had hoped.

Another way to make choices is to look at frequency in large general English corpora, such as the British National Corpus (BNC) (`http://www.natcorp.ox.ac.uk/`) or one of the newspaper article corpora distributed by the Linguistic Data Consortium. Such corpora play a prominent role in much current research in Natural Language Processing.

For example, the average length of sentences in the BNC is 16 words. Hence we could base aggregation decisions on sentence length; for example we could say that two pieces of information should be aggregated and expressed in one sentence if and only if this aggregation brings average sentence length closer to 16 words/sentence. Of course aggregation decisions must consider other factors as well, such as semantic compatibility (for example, *John bought a radio and Sam bought a TV* is better than *John bought a radio and Sam bought an apple*).

A perhaps more basic problem is that rules based on a corpus which combines many types of texts intended for many audiences, such as the BNC, may not be appropriate for the context in which a specific NLG system is used. For example, because SkillSum users are likely to have below-average literacy skills, they should probably get shorter sentences than is the norm; indeed SkillSum sentences on average are only 10 words long.

Another problem with relying on a general corpus such as the BNC is that in many contexts there are strong conventions about choices, and these should be respected. For example, one version of SkillSum generated reports for teachers instead of for the people actually taking the test, and this version referred to test subjects as *learner*, because this is the standard term used by adult literacy tutors to refer to the people they are teaching. The perhaps more obvious word *student* is much more common in the BNC (it occurs 16 times more often than *learner*), and probably would be used in texts which used choice rules based on BNC frequency; but this would be a mistake, because teachers in this area have a strong convention of using the word *learner* instead of the word *student*.

Hence a better alternative is to try to imitate the choices made in a corpus of human-authored texts which are intended to be used in the same context as the texts we are trying to generate. This can be done in two ways: we can either collect a corpus of texts written by many authors which are representative of human-authored texts in this domain, or we can collect a corpus of texts from a single author, perhaps someone we believe is a particularly effective writer. In other words, we can try to imitate the **style** of texts in the genre as a whole, or the **style** of a particular individual author.

Yet another approach to making microplanning choices is to allow the reader to directly control these choices. In practice this seems most successful if choices are presented to the user as **stylistic** ones, such as level of formality.

These approaches are summarised in Table 1.

4 Style 1: Explicit Stylistic Control

Perhaps the most obvious solution to the choice problem is to directly ask users what choices they prefer in texts generated for them. After all, software which presents information graphically usually gives users many customisation options (colours, fonts, layout, etc), so why not similarly give users customisation options for linguistic presentations of information?

It is not feasible to ask users to directly specify microplanning choice rules, because there are too many of them; for example, SkillSum has hundreds of different constraints, and its preference functions contain dozens of components. Hence users are usually asked to specify a few high-level parameters which the NLG system then maps into the actual low-level microplanning choice rules. For example, rather than directly specify aggregation rules, a SkillSum user could specify a preferred average sentence length (either numerically or via a linguistic term such as *short*, *medium*, or *long*). This length could be used by the aggregation system as described above (Section 3). Similarly, rather than specify specific lexical choice rules for individual concepts, the user could specify whether he wants informal, moderately formal, or very formal

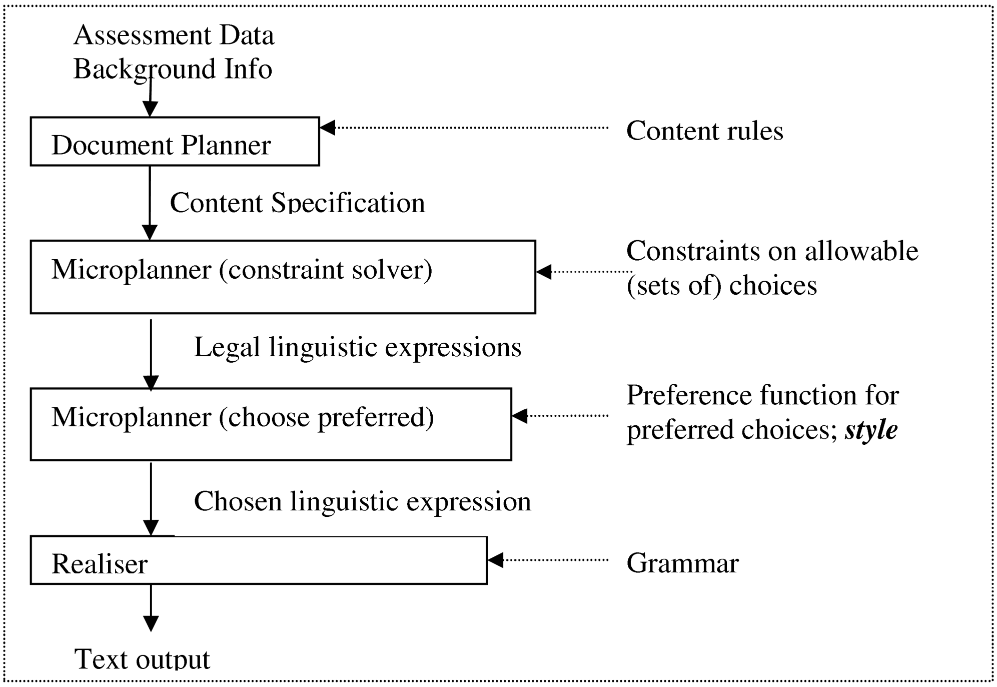

Assessment Data
Background Info
Document Planner
Content rules
Content Specification
Microplanner (constraint solver)
Constraints on allowable (sets of) choices
Legal linguistic expressions
Microplanner (choose preferred)
Preference function for preferred choices; ***style***
Chosen linguistic expression
Realiser
Grammar
Text output

Figure 3. SkillSum architecture

Explicit Control	Allow the user to specify the choices that she prefers. Choices are usually presented as stylistic ones.
Conform to Genre	Imitate the choices made in a corpus of genre texts.
Imitate Individual	Imitate the choices made by an individual writer.

Table 1. Three ways of using style to control choices in NLG

language; whether he prefers common words with many meanings (such as *got*) or less common words with fewer meanings (such as *answered*); and so forth. These general preferences could then be examined by SKILLSUM's detailed lexical choice rules. Such general preferences are usually perceived by users as *stylistic* preferences.

Although some of SKILLSUM's internal choice rules did refer to general preferences such as frequency vs. number of meanings, SKILLSUM users were not allowed to directly control these. Instead, the SKILLSUM developers refined the rules and preferences based on feedback and suggestions from literacy teachers and students. In other words, users made change requests to a developer instead of directly controlling the system. This is not ideal, but it means we did not have to deal with the difficult problems of designing an appropriate user interface for soliciting preferences [24] and also ensuring that SKILLSUM was robust enough to generate appropriate texts for any preference settings, no matter how bizarre.

Other NLG projects have tried to explicitly allow users to specify high-level stylistic preferences. For example, WebbeDoc [7] allowed users to specify level of formality, amount of technical content and vocabulary, literacy level, 'coolness', and role (e.g., doctor or patient); the system then generated a text according to these stylistic parameters. WebbeDoc did this using a 'Master Document' which encoded a rich representation of information that could be communicated in a document, and ways this information could be expressed; WebbeDoc then selected appropriate pieces of the Master Document, based on the stylistic settings, and combined these into a generated text. WebbeDoc's master document had to be carefully designed so that the above combination strategy did not result in incoherent texts. Perhaps the main long-term challenge in this approach is developing techniques, especially at the microplanning level, for automatically integrating master document segments into coherent texts. This may require changing some of these texts at the microplanning level, for example to ensure that appropriate referring expressions are used.

Another approach was taken by Paiva and Evans [16], who tried to base their controls on statistical analyses of texts. They analysed the surface linguistic features in a corpus of texts, and used factor analysis to cluster these on two dimensions. Their first dimension seemed to capture whether texts involved the reader or were distant from the reader; the second focused on the type of references used (e.g., pronouns or full noun phrases). In other words, although the dimensions were produced by factor analysis, they seemed to capture some notion of what humans would call style. Paiva and Evans' analysis was inspired by Biber's analysis [4], although they used fewer dimensions (essentially because they were working with texts in a limited genre). Paiva and Evans then built an NLG system which could produce texts with user-specified values of their two dimensions. This system was based on a model of how individual microplanner choices affected these dimensions; this model was created by statistically analysing the ratings (in the two dimensions) of texts generated with random microplanner choices.

In short, while the WebbeDoc developers choose intuitively appealing stylistic dimensions and explicitly coded how these dimensions affected the generation process, Paiva and Evans used statistical techniques to derive both the dimensions and the rules that linked dimensions to actual generation decisions.

While both of the above systems are very interesting, it perhaps is worth pointing out that both have only been demonstrated to work on a small set of examples. It seems likely that there would be major engineering challenges in scaling either system up so that it could robustly generate large numbers of varied texts.

Another constraint-based NLG system, ICONOCLAST [18], enabled low-level style preferences such as paragraph length, sentence length, word length, technical terms, passive voice and graphical impact to be configured by manipulating sliders in a graphical user interface (www.itri.brighton.ac.uk/projects/iconoclast/walk/trial.html). A user's selections did not change the constraints directly, but instead the selections modified weights associated with violating soft constraints. These in turn were used to compute a cost function associated with each output text. Allowing soft constraints to be violated with varying costs offers one solution to the problem of introducing user preferences into constraint satisfaction problem solving. It is unfortunate that the ICONOCLAST style interface has not yet been evaluated with users, because there remains the problem of whether it is reasonable to expect users to understand how to choose sets of low-level style parameters. Intuitively, making such low-level language choices seems a difficult task. ICONOCLAST's Web interface also grouped low-level style preferences into higher-level style profiles such as, "broadsheet" and "tabloid", an approach that looks more promising in terms of usability. Indeed, determining suitable style profiles and evaluating their usability would be fruitful topics for future research.

An obvious source of existing style profiles would be the in-house style guidelines used by newspaper copy editors. Some newspapers publish their own style guidelines, e.g., the Guardian (www.guardian.co.uk/styleguide), however, these tend to have rather vague directives such as "vary sentence length" which would be hard to encode. What they might offer, though, are concrete lists of style features that these publications consider to be important. A comparative study of in-house style guidelines from different publishers would show whether publishers vary, and how.

5 Style 2: Conform to a Genre

Another approach to making choices is to imitate a corpus of human-written texts. As mentioned above, imitating a general corpus such as the BNC is problematical because it ignores constraints due to the domain, the genre, and the characteristics of the user population; these are very important in many NLG applications. However, we can try to imitate a corpus of texts written for the NLG system's application, domain, and users. In other words, we can analyse a corpus of human-written texts in the genre; learn the words, syntactic structures, and so forth that human writers use; and program our NLG system to imitate these choices.

This imitation can be done in a number of different ways. In particular, we can manually analyse the corpus and extract choice rules from it; we can automatically extract choice rules using statistical corpus analysis and statistical generation techniques; or we can use a combination of these techniques.

For example, when building SKILLSUM we collected a small corpus of 18 example human-written reports; these were written by two tutors (one of which specialised in literacy and one of which specialised in numeracy). We analysed this, mostly by hand (since the corpus was quite small), primarily to create hard constraints for SKILLSUM's microplanner. In other words, we tried to get SKILLSUM to generate appropriate genre texts by only allowing it to make choices which we observed in the corpus.

To take a concrete example, the corpus texts used the verbs *scored*, *answered*, and *got* (e.g., *you answered 20 questions correctly*); but they did not use the verbs *responded* (e.g., *you responded to 20 questions correctly*) or *aced* (e.g., *you aced 20 questions*). Hence a hard constraint on SKILLSUM is that it should not use *responded* or *aced*. In a sense, this suggests that SKILLSUM reports should be moderately formal; and if style was being explicitly specified as in Section 4, then this level of formality might be explicitly specified. But in the genre-corpus approach we don't specify such high-level stylistic parameters such as level of formality, instead we directly specify low-level choices such as which verbs can be used when communicating numerical performance on an assessment.

In a few cases we allowed SKILLSUM to deviate from the corpus; but this often proved ill-advised. For example, we programmed SKILLSUM to use *right* instead of *correct* or *correctly*, for example *you got 20 questions right* instead of *you got 20 questions correct*. We did this because *right* is much more common in the BNC, and hence we thought it would be easier to read. Although the tutors agreed that *right* could be used, when we asked 25 students enrolled in a literacy course about this choice, 23 (92%) preferred *correct* over *right*, and 24 (96%) preferred *correctly* over *right*. This suggests that allowing SKILLSUM to use a word which was not in the corpus, at least in this example, was a mistake.

Of course the SKILLSUM microplanner needs a preference function (to choose between allowable options) as well as hard constraints (to say which options should be considered). In theory preferences between choices can be specified by looking at frequencies, but this is more controversial. For example, in the SKILLSUM corpus *scored* is more common than *answered* or *got*, so *scored* should be preferred under a pure frequency-based metric. However, frequencies are not always a good guide [22], because they may reflect the writing habits and preferences of a few individual corpus authors. In fact, *scored* was only used in reports written by one tutor, but it has the highest frequency because this tutor contributed the most texts to the SkillSum corpus. Hence in this case corpus frequency is really telling us about the linguistic preferences of the biggest contributor to the corpus; as we have no a priori reason to believe that this person is a better writer than the other corpus contributor, we need to interpret corpus frequency with caution.

In terms of methodology, SkillSum's rules were based on manual inspection of the corpus. Another possibility is to use machine learning techniques to automatically create rules or decision trees from a corpus; these can then be manually inspected by developers, who can modify the rules if necessary. This approach was used in SumTime-Mousam [22], which generated weather forecasts. SumTime-Mousam's microplanning rules (which focused on lexical choice, aggregation, and ellipsis) were based on careful analysis of a corpus of human-authored weather forecasts. Although most of these analyses were initially done using machine learning or statistical techniques, the rules suggested by the analyses were examined by developers and discussed with domain experts before they were added to the system [23]. This was especially important in cases where the corpus analysis showed that there was considerable variation in how different individuals made a choice. An evaluation with forecast users showed that the texts produced by SumTime-Mousam were very good, indeed in some cases they were perceived as being better than the human-written texts in the corpus.

Genre-specific microplanning rules can also be produced purely by machine learning and statistical analysis techniques, without having rules inspected by human developers or domain experts. This approach was used by Belz [1], who reimplemented some of SumTime-Mousam's functionality using a pure learning approach. An obvious advantage of this approach is that it is cheaper, since less human input is needed. Another advantage is that the rules do not have to be understandable by humans, as is the case with SumTime-Mousam's semi-automatic approach. However, a disadvantage is that developers, domain experts, and users cannot suggest that rules be modified based on their experience. An evaluation that compared Belz's system, SumTime-Mousam, and the human-written corpus texts [2] suggested that SumTime-Mousam's texts were on the whole better than Belz's texts, but Belz's texts were still quite good and in particular were sometimes better than the human-written corpus texts.

Perhaps the biggest problem we have faced in using machine learning techniques (whether semi-automatic or fully automatic) to learn microplanning choices in our nlg projects is obtaining a sufficiently large corpus. Although a few nlg systems such as SumTime-Mousam generate texts which are currently written by humans, it is more common for nlg systems to generate texts which are not currently manually written. In such cases it is not possible to get large corpora of naturally-occurring texts. In principle, one could analyse the microplanning choices made in related naturally-occurring texts, but this would require knowing which microplanning choices observed in the related texts could be applied to the nlg texts, and which could not.

In the SkillSum context, for example, domain experts (tutors) do not currently write reports about the results of assessments, instead they orally discuss results with their students. We could in principle obtain a corpus of transcripts of discussions about assessments between tutors and students, and use learning and statistical techniques to analyse the choices made in the transcripts. But this is of limited utility unless we know which microplanning choices observed in the oral transcripts are also appropriate for written reports (lexical choice?), and which are not (aggregation?).

In other words, it would be much easier to use machine learning techniques to learn microplanning choices if we had a good understanding of which choices were stable across 'sub-styles' in a genre and which were not. Unfortunately, little currently seems to be known about this topic.

6 Style 3: Imitate a Person

A final style-related approach to making linguistic decisions is to imitate a person. As mentioned above, one problem with imitating a multi-author corpus is that different authors have different preferences between choices (in other words, different styles). Hence the frequencies in a corpus may reflect the choices of only a few authors who contributed the most texts rather than the best choice, as mentioned above. Also, choosing the most frequent choice every time may lead to inconsistencies which users dislike, essentially because this mixes the style of multiple authors [23].

An alternative is to try to imitate the linguistic choices made by a single person, perhaps someone who is known to be an effective writer in this genre. Imitating a single person increases consistency between choices, and also is likely to increase choice effectiveness if this person is an exceptionally good writer. However, a corpus from one individual is likely to be smaller and have worse linguistic coverage than a corpus with contributions from many people. Also, very good writers are likely to be very busy, which can make it difficult to discuss things directly with them.

SkillSum partially followed this approach when making decisions about content. More precisely, SkillSum generates two kinds of reports, literacy and numeracy, and the SkillSum corpus contains reports from two authors, one of whom is a literacy expert and the other of whom is a numeracy expert. When making some high-level decisions about the content of SkillSum's literacy reports, we tended to favour the choices make in the texts written by the literacy tutor; similar we focused on the numeracy tutor's choices when making choices about SkillSum's numeracy reports.

McKeown, Kukich. and Shaw [14] used this approach when building PlanDoc, an nlg system which produced summaries of the results of a simulation of changes to a telephone network. They interviewed a number of people to establish the general requirements of PlanDoc, but they asked a single very experienced domain expert to write all of the texts in their corpus. They do not give details of how they analysed and used the corpus, but it seems to have been a manual analysis rather than one based on learning or statistical techniques.

Another approach to individual style is to try to imitate the style of the *reader*, that is to generate texts in the style that the reader prefers when reading texts in this genre. Different people have different preferences. For example people who are very poor readers may do best with very short (5 word) sentences, people who are moderately poor readers may prefer 10 word sentences, people with average skills may prefer 15-20 word sentences, etc. We could directly ask people about their preferences, as discussed in Section 4. However this approach is limited in that most people will probably only be willing to explicitly specify a small number of preferences.

Perhaps the most advanced work in this area is that of Walker and her colleagues [28]. They asked users to explicitly rate 600 texts generated by their NLG system with random microplanning choices. They employed learning techniques to determine which sets of microplanning choices produced texts preferred by each user, and from this created choice models for each user, which could be loaded into the microplanner. Their experiments suggested that users did indeed prefer texts generated using their personal choice models. Walker *et al* also commented that they believed reasonable individual choice models could be extracted from ratings of 120 texts, and getting this number of ratings is probably more realistic than getting 600 ratings from each user.

Walker *et al* did not really consider lexical choice, which is a shame because we know that there are substantial differences in the meanings that different individuals associate with words [21]. This has been reported in many contexts, including weather forecasts [22], descriptions of side effects of medication [3], and interpretation of surveys [26]. It was also an issue in SkillSum. For example, while developing SkillSum we asked 25 people enrolled in a literacy course to tell us what kind of mistake was in the sentence

I like apple's

72% said this was a *punctuation* mistake but 16% said this was a *grammar mistake* (the rest didn't think there was anything wrong with this sentence). Hence if we want to tell someone that he or she has problems with this kind of construct, we should probably refer to it as *grammar mistake* for the first group, and *punctuation mistake* for the second. Note that while this may sound like a small point, in fact some SkillSum users got quite annoyed when SkillSum told them they were bad at something which they thought they were good at. For example, if a SkillSum user made the above mistake and SkillSum told him he had problems with *punctuation*, the user might get annoyed if he interpreted *punctuation* to just mean commas and full stops (periods), since he had not made any mistakes with these.

Perhaps the key problem in doing this kind of tailoring is getting sufficient data about the individual; how do we actually find out how he or she uses words? If we only need data about a small number of lexical choices, that we could use an approach similar to Walker *et al*; but this is unlikely to be feasible if we need information about many different lexical choices.

An alternative approach might be to analyse a large corpus of texts that the user has written, on the assumption that the style used in texts the user writes is similar to the style preferred by the user in texts that he or she reads. Lin [12] looked at one aspect of this in her investigation of distributional similarities of verbs in a corpus of cookery writing to find alternatives for how different recipe authors expressed the same concept (e.g., "*roast* the meat in the oven" vs. "*cook* the meat in the oven"). To the best of our knowledge larger-scale investigations of larger sets of style choices have not yet been tried; one concern is that many people (including most SkillSum users) do not write much, which would make it difficult to collect a reasonable corpus of their writings.

Data-scarcity becomes an even larger problem if we want to create models of individual linguistic preferences in specific genres. Ideally we would like not just a fixed set of linguistic preferences for a particular individual, but rather a mechanism for creating preference rules that express how text should be written for a particular individual in a specific genre. Again we are not aware of any existing research on this issue.

NLG might also employ research from the area of text categorisation and author identification, e.g., Stamatatos *et al.* [27], that attempt to identify authorship using machine learning of word or character n-grams from a corpus. Being very speculative, if an NLG system could generate outputs with different combinations of microplanning parameters, perhaps an authorship identification system could be used to select the text which was most similar to the target author. Of course there are many issues that would need to be resolved before this could be done, not least of which is that existing author identification systems work with human-generated texts, not computer-generated texts.

7 Research Issues

As should be clear from the above, there are numerous research issues in this area that can be explored, for both technological reasons (building better NLG systems) and scientific reasons (enhancing our understanding of style). A *few* of these challenges are:

- *Explicit stylistic controls:* What stylistic controls make sense to human users, and how can these be 'translated' into the very detailed choices and preferences that control NLG microplanners?
- *Conform to a genre:* How are rules derived from a genre corpus most likely to differ from rules derived from a general corpus? In other words, how do genre texts actually differ from non-genre texts? Are there rules which are unlikely to vary, and hence could be derived from a general corpus?
- *Individual stylistic models:* How can we get good data about an individual's language usage and preferences? What aspects of language usage are most likely to vary between individuals? How can we combine a (non-user-specific) genre language model with a (non-genre specific) individual language model?
- *What is the impact of style:* Generated texts can be evaluated in many different ways, including preference (e.g., do people like a text), readability (e.g., how long does it take to read a text), comprehension (e.g., how well do people understand a text), and task effectiveness (e.g., how well does a text help a user to do something). Which of these measures is most (and least) affected by adding stylistic information to an NLG system?

To conclude, we believe that style is an important aspect of generating effective and high-quality texts, and we are very pleased to see that an increasing number of NLG researchers are investigating style-related issues. We hope this research will lead to both better NLG systems, and also to a deeper scientific understanding of style in language.

Acknowledgements

We would like to thank our colleagues in Aberdeen and Milton Keynes, the anonymous reviewers, and the tutors we worked with in skillsum for their insightful comments and suggestions. This work was funded by PACCIT-LINK grant ESRC RES-328-25-0026.

REFERENCES

[1] Anja Belz, 'Statistical generation: Three methods compared and evaluated', in *Proceedings of ENLG-2005*, pp. 15–23, (2005).

[2] Anja Belz and Ehud Reiter, 'Comparing automatic and human evaluation of NLG systems', in *Proceedings of EACL-2006*, pp. 313–320, (2006).

[3] Dianne Berry, Peter Knapp, and Theo Raynor, 'Is 15 per cent very common? informing people about the risks of medication side effects', *International Journal of Pharmacy Practice*, **10**, 145–151, (2002).

[4] Douglas Biber, *Variation across speech and writing*, Cambridge University Press, 1988.

[5] Bruce Buchanan, Johanna Moore, Diana Forsythe, Guiseppe Carenini, Stellan Ohlsson, and Gordon Banks, 'An interactive system for delivering individualized information to patients', *Artificial Intelligence in Medicine*, **7**, 117–154, (1995).

[6] Alison Cawsey, Ray Jones, and Janne Pearson, 'The evaluation of a personalised health information system for patients with cancer', *User Modelling and User-Adapted Interaction*, **10**, 47–72, (2000).

[7] Chrysanne DiMarco, Graeme Hirst, and Eduard H. Hovy, 'Generation by selection and repair as a method for adapting text for the individual reader', in *Proceedings of the Workshop on Flexible Hypertext, 8th ACM International Hypertext Conference*, (1997).

[8] Eli Goldberg, Norbert Driedger, and Richard Kittredge, 'Using natural-language processing to produce weather forecasts', *IEEE Expert*, **9**(2), 45–53, (1994).

[9] Trevor Harley, *The Psychology of Language*, Psychology Press, second edn., 2001.

[10] Xiaorong Huang and Armin Fiedler, 'Proof verbalization as an application of NLG', in *Proceedings of IJCAI-1997*, volume 2, pp. 965–972, (1997).

[11] Benoit Lavoie and Owen Rambow, 'A fast and portable realizer for text generation', in *Proceedings of the Fifth Conference on Applied Natural-Language Processing (ANLP-1997)*, pp. 265–268, (1997).

[12] Jing Lin, 'Using distributional similarity to identify individual verb choice', in *Proceedings of the Fourth International Natural Language Generation Conference,*, pp. 33–40, (2006).

[13] Kathleen McKeown, *Text Generation*, Cambridge University Press, 1985.

[14] Kathleen McKeown, Karen Kukich, and James Shaw, 'Practical issues in automatic document generation', in *Proceedings of ANLP-1994*, pp. 7–14, (1994).

[15] Mick O'Donnell, Chris Mellish, Jon Oberlander, and Alistair Knott, 'ILEX: an architecture for a dynamic hypertext generation system', *Natural Language Engineering*, **7**, 225–250, (2001).

[16] Daniel Paiva and Roger Evans, 'Empirically-based control of natural language generation', in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 58–65, Ann Arbor, Michigan, (June 2005). Association for Computational Linguistics.

[17] Paul Piwek, Richard Power, Donia Scott, and Kees van Deemter, 'Generating multimedia presentations: From plain text to screenplay', in *Intelligent Multimodal Information Presentation*, Kluwer, (2005).

[18] Richard Power, Donia Scott, and Nadjet Bouayad-Agha, 'Generating texts with style', in *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'03)*, pp. 444–452, (2003).

[19] Ehud Reiter and Robert Dale, *Building Natural Language Generation Systems*, Cambridge University Press, 2000.

[20] Ehud Reiter, Roma Robertson, and Liesl Osman, 'Lessons from a failure: Generating tailored smoking cessation letters', *Artificial Intelligence*, **144**, 41–58, (2003).

[21] Ehud Reiter and Somayajulu Sripada, 'Human variation and lexical choice', *Computational Linguistics*, **28**, 545–553, (2002).

[22] Ehud Reiter, Somayajulu Sripada, Jim Hunter, and Jin Yu, 'Choosing words in computer-generated weather forecasts', *Artificial Intelligence*, **167**, 137–169, (2005).

[23] Ehud Reiter, Somayajulu Sripada, and Roma Robertson, 'Acquiring correct knowledge for natural language generation', *Journal of Artificial Intelligence Research*, **18**, 491–516, (2003).

[24] Ehud Reiter, Somayajulu Sripada, and Sandra Williams, 'Acquiring and using limited user models in NLG', in *Proceedings of the 2003 European Workshop on Natural Language Generation*, (2003). Forthcoming.

[25] Ehud Reiter, Sandra Williams, and Leslie Crichton, 'Generating feedback reports for adults taking basic skills tests', in *Applications and Innovations in Intelligent Systems XIII: Proceedings of AI-2005*, pp. 50–63. Springer, (2005).

[26] Michael Schober, Frederick Conrad, and Scott Fricker, 'Misunderstanding standardized language in research interviews', *Applied Cognitive Psychology*, **18**, 169188, (2004).

[27] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis, 'Automatic text categorization in terms of genre and author', *Computational Linguistics*, **26**, 471–495, (2000).

[28] Marilyn Walker, Amanda Stent, François Mairesse, and Rashi Prasad, 'Individual and domain adaptation in sentence planning for dialogue', *Journal of Artificial Intelligence Research*, **30**, 413–456, (2007).

[29] Sandra Williams and Ehud Reiter, 'Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application', in *Proceedings of Corpus Linguistics workshop on using Corpora for NLG*, (2005).

[30] Sandra Williams and Ehud Reiter, 'Generating readable texts for readers with low basic skills', in *Proceedings of ENLG-2005*, pp. 140–147, (2005).

[31] Sandra Williams and Ehud Reiter, 'Generating basic skills reports for low-skilled readers', *Natural Language Engineering*, (in press).