_lognostics
tools for vocabulary research

Lexical Signatures in foreign language free-form texts.

Paul Meara, Gabriel Jacobs and Catherine Rodgers

University of Wales Swansea

Abstract

This paper presents an investigation into the extent to which the lexical choices made by learners of a second language (L2) are distinctive. It follows on from an earlier paper by the same authors in which a neural network was successfully trained to mark a set of texts produced by L2 learners to the same standard, within broad categories, as had been awarded by experienced human markers. For this present paper, we examined a set of L2 texts and searched them for unique lexical choices ('lexical signatures'). The results suggest a possible explanation for the success of the neural-network trial, and may have some practical implications for determining the levels of achievement reached by L2 learners.

Introduction

This paper is concerned with the idea of 'lexical signatures'. When we write, whether in a first or second language (L1 or L2), each of us makes lexical choices, and these choices reflect our ability to operate effectively in a language. The choices we make tell readers or listeners about ourselves, and provide them with clues about our language skills. Skilled readers can easily recognise a literary author's particular style, for example, or whether a text is intended for children. Several factors contribute to style, but a very large factor is the author's choice of particular words.

A great deal of work has been carried out looking at the way writers choose words, and how their style can be characterised in lexical terms (see for example, HOLMES 1994 for an overview of the question of authorship attribution.) It is normally assumed, however, that this type of analysis can only be applied to very advanced writers – typically novelists, or poets. Our interests lie in an area which has not been studied in this way before: the lexical choices that L2 speakers make in ordinary tasks. In particular, this paper is concerned with how far and how easily lexical choices (lexical signatures) can be identified in texts written by L2 learners.

In some of our previous work, we have been struck by the diversity of lexical choice made by L2 learners. For example, in a collection of 77 short letters written as part of a low-level English language examination, we counted a total of 1,270 different words, of which only 3 occurred in all the texts. This was despite the fact that the contents of the letters were very highly constrained by the examination question, which simply asked the writers to confirm arrangements for a visit to a factory. This finding suggests that even at relatively low levels of proficiency, L2 learners are far from uniform in their lexical choices.

The question we are interested in here is how easily we can identify lexical signatures in L2 writers. Obviously, at the most basic level, texts produced by L2 writers differ from each other, and some texts contain words which are not included in other texts. In fact, given any set of free-form texts, it is almost always possible to find some words which occur only in one text. In our examination texts, for example, more than half of the 1,270 words in the entire corpus occurred in only one text: obviously, these words uniquely identify a text. In a way, however, this is not a particularly interesting finding. A great deal of work is required to identify these words, and in the final analysis, there is not much to be said about them other than that they are used by only one writer. A more interesting question is whether we can find **patterns of lexical choice** among words which occur more frequently in a set of texts. These words are much easier to identify, and because they tend to be more common words, they appear to be less serendipitous than the words which occur in only one text.

The work that follows is based on a set of 59 essays in French, produced by non-native-speaking learners of French, undergraduates at a British university. For reasons which will become apparent later, we originally planned to work with 64 texts, but five were lost due to reasons beyond our control, and we were unable to replace these missing texts from our student cohort. The essays were based on a task which requires the students to describe and analyse a picture produced by the Scottish Tourist Board, which had appeared in the French press. The picture shows a framed, sepia-coloured photograph of a young man and a woman holding hands in a remote Scottish glen, while the text, tastefully set against a tartan background, exhorts people to spend their holidays in Scotland, and provides details of where to apply for a brochure. This task tends to produce unimaginative responses, which are very similar in content and style.

How can we describe the lexical choices in texts of this sort in an interesting way? The methodology we have developed is basically very simple. First we choose a small set of target words which we then use to describe each text. For example, suppose we decide that we are interested in the words:

page, *blanc*, *couleur*, *pas*, *office*, *comme*. word list A

Now, let us suppose that some of these words appear in text X while others do not. We can now derive a binary part-description of text X in which words that occur are represented as 1s, while missing words are represented by 0s. In this way, the description of text X with respect to word list A might be:

1 1 1 0 0 0

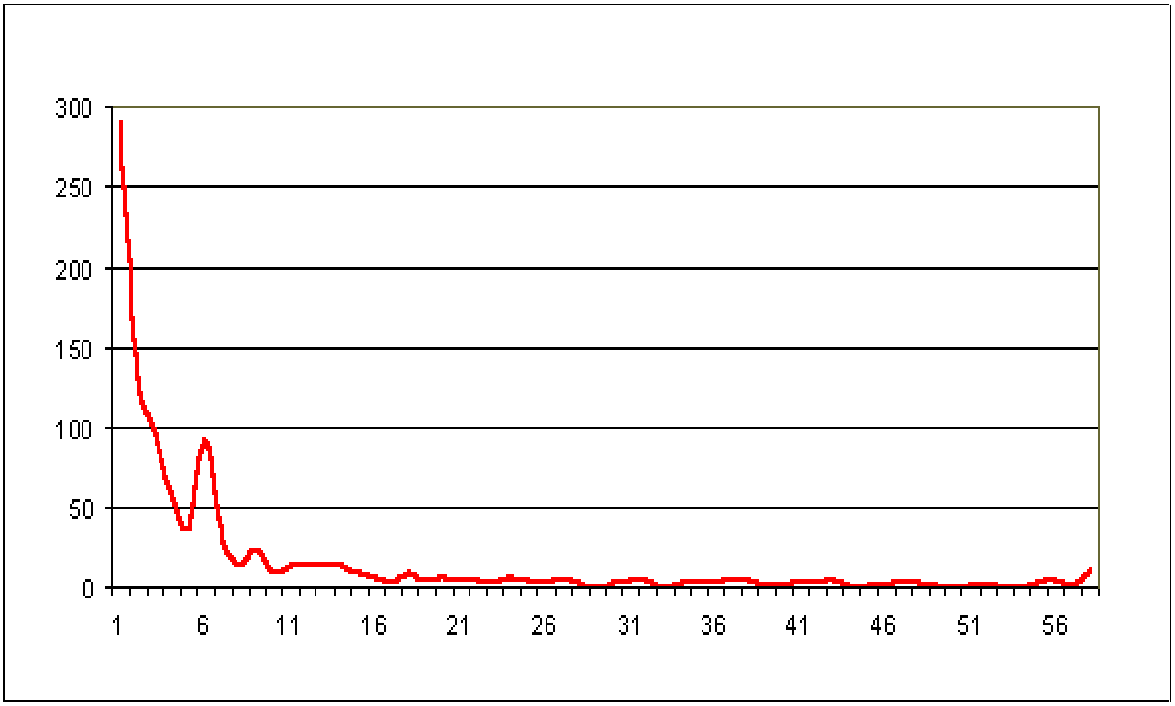
We read this as showing that target words 1 to 3 occur in the text, while target words 4, 5 and 6 do not.

This notation is very compact, and it allows us to carry out a number of simple comparisons between sets of texts in an efficient way. Specifically, the notation offers us a straightforward method of exploring the idea that L2 writers make unique lexical choices, even in unpromising and undemanding tasks.

Method

The 59 texts produced by the students were first assembled into a single corpus. All the words that appeared in the corpus were listed (a total of 2,272 unlemmatised words), and the number of texts containing each word was catalogued. These data are summarised in Figure 1, which shows the number of different words occurring in a single text, the number of different words occurring in two of the texts, the number of words occurring in three of the texts, and so on up to the number of words appearing in all 59 texts. Figure 1 clearly indicates that most words occur in a small number of texts, suggesting that there is relatively little lexical overlap between texts.

Figure 1: The number of different words appearing in N texts



We can now test the claim that L2 writers tend to make unique lexical choices by taking a set of N words from this list, and checking their occurrence in the texts. How big should N be? Given 59 texts, the strongest possible test of our conjecture would be to take a set of only six words. Six words gives us 2^6 = 64 possible patterns of choice: that is, given six words which can either be present in a text or not, there are a total of 64 possible patterns (ranging from 0 0 0 0 0 0 to 1 1 1 1 1 1 in binary notation). Our conjecture that L2 learners tend to produce unique lexical signatures suggests that we ought to find a very large number of the 64 possible combinations among our data.

Table 1: three sets of randomly selected words. Set H = words occurring in many texts; set M = words occurring in about half the texts; set L= words occurring in few of the texts

Set H: words appearing in a large number of texts
sur réalité nessie charme peut cette plus sont avec aussi se son image homme
entre blanc page par ce vous femme très tourisme ensorcelant office guide jaune
couleurs mot être paysage printemps comme mais mots elle deux gens tout rattrapera
pas photo pays noir dessous réveille tartan bas lettres ils

Set M: words occurring in about half the texts
bon montagnes gratuit partie adresse couleur vacances fond bien petit même centre
remplir ou bleu impression aide rouge questions donne chapeau monstre couple nom
grande petite haut sa beaucoup vert voit temps lac nous jupe où peu encore phrase
si internet ont trouve gauche assez porte but romantique veut

Set L: words occurring in few texts
vêtements petits facile grandes beau trouver ceux nature air lecteur idée fois blanches
faut beauté quand choses mystère traditionnel vieux tous grand documentation pense va
titre vie sens doit leur peinture traditionnelle côté non plein aider utilisé style
quelques écriture tradition livre cela donc lire puis soleil allé quelque je

We tested this idea in the following way. First we selected three sets of 50 words. Set H (= High Frequency) were 50 words that occurred in many of the texts, specifically between 32 and 50 texts; set M (= Middle Frequency) were 50 words which occurred in about half the texts, specifically between 19 and 29 texts; set L (= Low Frequency) were 50 words which occurred in only a few texts, specifically between 10 and 13 texts. Fifty words was an arbitrary choice, large enough to produce interesting data but small enough to be manageable. These word sets are listed in Table 1. It is important to note that most of the words are common French terms. We are not dealing with arcane items here: the few unusual words such as *tartan* and *Nessie* reflect the specific content of the advertisement.

From each set of 50 words, we developed 25 randomly selected subsets, each subset consisting of 6 words (clearly, there is some overlap between the subsets). This gave us a set of 75 subsets, examples of which are listed in Table 2. Twenty-five of the subsets come from set H, 25 from set M and 25 from set L.

We can now ask two questions of these data:

1. How many distinct data patterns (distinct signatures) can be found?
2. How many of these distinct signatures are unique?

A distinct signature is one which is different from all the other signatures in at least one place. Not all distinct signatures are unique: a small number of the signatures are produced by more than one writer.

The answers to the above questions for the data in Table 3 turned out to be 32 distinct patterns, of which 21 are unique. But, clearly, a better overall picture can be obtained by averaging out the figures for a large number of different subsets. We therefore repeated this type of analysis for each of the 25 subsets from set H, set M and set L, making a total of 75

Table 2: examples of the six-word target subsets

target subsets from wordset H

ils: lettres: par: elle: femme: peut:
noir: office: nessie: sont: page: homme:
se: noir: tourisme: bas: plus: son:
photo: réveille: comme: ensorcelant: femme: rattrapera:
noir: page: printemps: photo: ensorcelant: deux:
page: blanc: couleurs: pas: office: comme:
par: page: lettres: dessous: pas: réveille:
gens: charme: lettres: femme: mot: guide:
blanc: photo: son: homme: ce: entre:
très: blanc: mots: dessous: jaune: charme:

target subsets from wordset M

petite: couleur: bien: nous: grande: ont:
montagnes: ont: bien: encore: grande: chapeau:
fond: voit: couple: centre: bleu: chapeau:
remplir: gratuit: temps: impression: centre: nous:
couple: questions: sa: chapeau: monstre: gauche:
assez: peu: encore: lac: donne: sa:
porte: voit: gauche: temps: jupe: questions:
voit: peu: chapeau: vert: grande: porte:
ou: où: monstre: couleur: rouge: chapeau:
veut: haut: bleu: si: couple: bon:

target subsets from wordset L

leur: choses: livre: beauté: quand: idée:
mystère: fois: choses: tradition: doit: je:
tradition: pense: soleil: tous: documentation: quand:
idée: côté: quelques: cela: documentation: donc:
donc: blanches: vieux: puis: traditionnel: peinture:
pense: mystère: cela: plein: écriture: va:
quand: traditionnelle: je: doit: beauté: air:
grand: petits: non: utilisé: style: donc:
doit: leur: beau: vie: traditionnelle: ceux:
traditionnelle: pense: sens: idée: plein: aider:

separate analyses. This process is summarised in Table 4, while Table 5 summarises the outcome.

Next, we generated binary descriptions of the texts, in terms of each of the subsets. This gave us a set of 59 binary descriptions for each of the 75 subsets. By way of an example, Table 3 shows the complete data for set H, subset 1.

Table 3: binary descriptions of 59 texts: subset H1

```
0 0 0 0 1 1 \rosab      0 1 1 1 0 1 \greni      1 0 1 1 1 1 \mitst
0 0 1 0 0 0 \coowe      0 1 1 1 1 0 \denge      1 0 1 1 1 1 \priap
0 0 1 0 0 1 \jonle      0 1 1 1 1 1 \hugri      1 0 1 1 1 1 \turd
0 0 1 0 1 0 \snepe      0 1 1 1 1 1 \mcgpa      1 1 0 0 1 0 \faich
0 0 1 1 0 0 \fiffe      0 1 1 1 1 1 \jamto      1 1 0 0 1 1 \jacet
0 0 1 1 0 1 \coopa      0 1 1 1 1 1 \bairh      1 1 0 1 1 1 \neada
0 0 1 1 0 1 \wiljo      0 1 1 1 1 1 \cosja      1 1 0 1 1 1 \tomed
0 0 1 1 1 1 \hucjo      1 0 0 0 0 1 \vivta      1 1 1 0 0 1 \heajo
0 0 1 1 1 1 \sorma      1 0 0 0 0 1 \ganio      1 1 1 0 1 0 \ricno
0 0 1 1 1 1 \davni      1 0 0 1 0 1 \kocmi      1 1 1 0 1 0 \paybe
0 0 1 1 1 1 \lewrh      1 0 0 1 1 1 \halra      1 1 1 0 1 0 \davbo
0 1 0 0 0 1 \allka      1 0 1 0 0 1 \wildo      1 1 1 0 1 0 \googl
0 1 0 0 1 1 \oxtgw      1 0 1 0 0 1 \brala      1 1 1 1 0 1 \samki
0 1 0 1 0 1 \colga      1 0 1 0 1 1 \palpi      1 1 1 0 1 1 \webja
0 1 0 1 1 1 \davan      1 0 1 0 1 1 \boxer      1 1 1 1 1 1 \lanpa
0 1 0 1 1 1 \thoan      1 0 1 1 0 0 \jonti      1 1 1 1 1 1 \unwbe
0 1 0 1 1 1 \mcdte      1 0 1 1 0 1 \colei      1 1 1 1 1 1 \kenma
0 1 0 1 1 1 \bebli      1 0 1 1 1 0 \waldo      1 1 1 1 1 1 \solpa
0 1 1 0 1 1 \halli      1 0 1 1 1 1 \decie      1 1 1 1 1 1 \carcl
0 1 1 0 1 1 \davwa      1 0 1 1 1 1 \groma
```

Table 4: summary of the data collection and analysis:

Assemble a small corpus of texts and enumerate the words in the corpus

Select 50 words occurring in most texts (set H)
Select 50 words occurring in about half the texts (set M)
Select 50 words occurring in few of the texts (set L)

Generate 25 subsets of 6 words from set H
Generate 25 subsets of 6 words from set M
Generate 25 subsets of 6 words from set L

Generate 75 binary descriptions of the texts, one for each subset

Table 5: mean distinct and mean unique signatures (59 texts, subset size = 6)

	distinct signatures	**unique signatures**	**ratio of unique to distinct signatures**
Set H	20.88	16.88	80.8
Set M	35.60	21.24	59.7
Set L	21.20	10.72	50.6

What do these results show us?

At the most superficial level, the data show that there is a some degree of variation among the texts as described by the subsets. On average, for set H, and for set L, approximately one third of the 64 possible binary combinations are actually used up, while for set M, the figure rises to half of the possible binary combinations. For set H, 80% of the combinations that are used are indeed unique. For sets M and L, the figures are lower – 59% and 50% of the patterns we find are unique. These average figures actually mask a great deal of variation, and this is summarised in Table 6, which shows the best and the worst figures produced by each subset at each of the frequency levels:

Table 6: performance of the best and worst subsets (59 texts, subset size = 6)

		Set H	**Set M**	**Set L**
different signatures	**best:**	34	41	25
	worst:	29	32	18
unique signatures	**best:**	21	29	18
	worst:	13	16	4

It can be seen from Table 6 that, even under exacting conditions (sets of only 6 words), a surprisingly large number of unique signatures can be identified. This suggests that if we relaxed our constraints, it might be possible to improve the number of unique signatures we find. In order to test this idea, we ran a new set of analyses, in which we used subsets 7, 8, 9 and 10 words (2^7 allows for 128 distinct binary patterns, 2^8 for 256 patterns, 2^9 for 512 patterns, and 2^{10} for 1,024 patterns). The results are summarised in Table 7.

Table 7: mean distinct and mean unique signatures for varying size subsets

# distinct signatures

subset size	**6**	**7**	**8**	**9**	**10**
set H	20.88	28.88	41.76	45.80	48.88
set M	35.60	43.56	49.72	54.40	55.12
set L	21.20	26.20	31.04	36.32	39.68

# unique signatures

subset size	**6**	**7**	**8**	**9**	**10**
set H	16.88	23.84	32.28	37.84	42.00
set M	21.24	31.80	42.28	50.48	51.80
set L	10.72	15.64	20.48	26.68	30.40

best performance at identifying unique signatures

subset size	**6**	**7**	**8**	**9**	**10**
set H	21	35	39	48	57
set M	29	41	51	7	57
set L	18	23	29	32	40

These data suggest that in some circumstances it is very nearly possible to select a set of 10 commonly occurring words that will uniquely identify a text. At this level, for sets H and M, the best subsets uniquely identify more than 90% of our cases. Some good subsets occur with smaller subsets from these sets. We suspect that a genetic algorithm approach might allow us to identify these very good discriminators with ease. Set L seems to be a less good source of target words, although even here, the worst subset managed to identify 35 distinct signatures, of which 28 – almost 80% - were unique.

Discussion

The analyses reported here arose from a discussion of some of our earlier work in which we suggested that lexical signatures might be used together with neural networks in order to evaluate the writing of non-native speakers (Meara, Rodgers and Jacobs 2000). In that work, we used sub-sets of 10 words, and found that we could almost always train a network to recognise the distinct signatures and assign them an appropriate mark. With hindsight, this result is not as astonishing as we thought it was. We suggested at the time that it was surprising to achieve this degree of reliability with the very small amount of information from the texts we were evaluating. However, one of our colleagues suggested that 2^{10} was such a large amount of information that our binary descriptions were in reality much richer than we thought. The analysis reported here seems to bear out this judgement. It may be that the neural network we used for that work was, above all, identifying levels of lexical uniqueness, and that this uniqueness correlated well with the subjective evaluative impressions of the human markers. That, in itself, in no way negates the interest or potential of the results, but they do need further investigation

In the meantime, we think we may have stumbled across a rather important finding, which may have implications in areas other than the one for which our work was intended. Presumably, very low-level learners do not show the degree of variation we have reported here, if only for the reason that their L2 vocabulary is more limited. This raises the question of whether unique lexical signatures could be used as a way of measuring progress in the learning of a second language. It seems plausible that less proficient learners, whose lexicons are limited in size, would tend to make similar lexical choices, and their lexical signatures would tend, therefore, not to be unique. On the other hand, learners with a good grasp of their second language, and a more extensive vocabulary, would be more likely to produce unique signatures. This suggests that it ought to be possible to compute a uniqueness index for learners, by calculating the number of times a small randomised lexical sample of their output matches samples found in a criterion group. We believe that this idea might provide an alternative approach to the problem of how written texts produced by non-native speakers could be evaluated in on-line settings.

References

Holmes, DI
Authorship attribution. *Computers and the Humanities*, 28,2(1994), 87-106.

Meara, PM C Rodgers and G Jacobs
Computational assessment of texts written by L2 speakers. *System,* 28(2000), 345-354.

Notes

This paper first appeared in *ITL Review of Applied Linguistics* 135-136(2002), 1-12.