Investigating style in a corpus of pharmaceutical leaflets: results of a factor analysis

Daniel S. Paiva
ITRI, University of Brighton
Lewes Road
BN2 4GJ, England,
Daniel.Paiva@itri.brighton.ac.uk

Abstract

In this paper we present an analysis of stylistic variation that uses a factor analytic technique to group the variables responsible for the bulk of the linguistic variation found in a corpus of pharmaceutical leaflets. Two main factors of variation were found and analysed in more detail; they also were compared with other two analyses.

1. Introduction

In this paper we are interested in characterising stylistic variation of texts with a possible application to natural language generation (NLG) — see Paiva (1999) for how we intend to use style in generation. To motivate our discussion, we start by introducing a scenario of pharmaceutical companies which needs to produce patient information leaflets for several of their medicines. Although some uniformity in the writing of a company would be desirable/expected (i.e., for it to have a 'company style'), some linguistic differences may be necessary for medicines aimed at specific public (e.g., child, elderly, etc.), or for medicines which are from different groups (e.g., analgesics, etc.) or which are taken by using a different vehicle (e.g., tablet, creams, etc.).

In this paper we show that this scenario is not unrealistic. We are dealing with a corpus of about 340 patient information leaflets that come from more than 40 companies. Our aim is to obtain the variables that represent the main stylistic variation occurring in the texts of our corpus.

We start by presenting examples which show how even leaflets from the same company can differ one from the other; the examples were taken from the section of our corpus about how a patient should take his medicine (see figures 1 and 2). The first pair of texts is from medicines Livial and Normegon, which are produced by *Organon*. Livial is a tablet for menopause symptoms relief (hormone replacement therapy), whereas Normegon is a solution for injection for treating infertility (for both men and women). The second pair is from medicines Efcortelan and Ceporex, produced by *Glaxo*. Efcortelan is a corticosteroid cream for treating skin disorders, and Ceporex is an antibacterial syrup used on infections.

We believe that these pairs of texts present differences in their style and this can be justified, among other things, by their different use of linguistic features. For example, the first instance in figure 1 uses more second person pronouns and imperative sentences whereas, in the second instance, the use of passive sentences and prepositions is more prominent. In figure 2, it is possible to say that the second instance uses more 'it' pronouns and is in a more verbal style than the first instance.

What is difficult to say, however, is that this kind of inspection can give a consistent account of the variation occurring along the texts in a corpus. In fact, with the amount of texts we are dealing with, the variation from one text to the other can be so subtle that manual inspections can often be misleading.

How do I take Livial? á**Livial**ñ

It is important to take this medicine only as directed by your doctor or pharmacist.

The recommended dose is one tablet every day.

Take your tablet at the same time each day.

Take a tablet marked with the corresponding day of the week. For example, if it is Wednesday, take the tablet marked Wednesday on the upper row of the pack. Follow the direction of the arrows and continue taking one tablet each day until the pack is empty.

You may start to feel better after a few weeks, but Livial may take up to 3 months to work fully.

How to take the tablets.

Swallow the tablets with some water or other drink. Do not chew the tablets.

⟨*TEXT TRUNCATED HERE*⟩

Using this medicine properly á**Normegon**ñ

Normegon should only be given by a doctor or nurse

How much: the dose is chosen by the doctor.

In female patients injections are given daily or sometimes every second day for about 10 days. In male patients injections are given several times a week for at least 10 - 12 weeks.

How to administer: the powder in one ampoule should first be dissolved in the fluid in the other ampoule. The injections should be given in muscle tissue (for instance in the buttock, upper leg or upper arm).

⟨*TEXT STOPS HERE*⟩

Figure 1: Section of medicines Livial and Normegon from *Organon***.**

How to use your cream á**Efcortelan**ñ

If your doctor has told you in detail how much to use and how often then keep to this advice.

If you are not sure then follow the advice on the back of this leaflet.

Unless told by your doctor:

- You should not use more than this.

- You should not use on large areas of the body for a long time (such as nearly every day for many weeks or months). Although Efcortelan Cream is generally regarded as safe even when used like this for many months it may be possible to produce side effects if overused. Such overuse may thin the skin so that it damages easily and some of the active ingredient may pass through the skin and affect other parts of the body, especially in infants and children.

⟨*TEXT TRUNCATED HERE*⟩

---

* Look at the label á**Ceporex**ñ

It should say who should take it, how many 5 ml spoonfuls and when. It should also give a date after which the medicine must not be used. If it does not or you are not sure, ask your doctor or pharmacist. (If prescribed for a child, make sure the medicine is taken as the label says).

* How to take your medicine

Use the 5 ml spoon that the pharmacist has given you to measure the amount of medicine to take.

It is best to take the syrup as it is. If you want you can add a little water to each dose just before you take it. Do not add water or other drinks to the bottle, that may stop it working properly.

⟨*TEXT TRUNCATED HERE*⟩

Figure 2: Section of medicines Efcortelan and Ceporex from *Glaxo*.

Factor 1	
Involved Production (+)	
Private verbs	.96
THAT deletion	.91
Contractions	.90
Present tense verbs	.86
Second person pronouns	.86
Analytic negation	.78
Demonstrative pronouns	.76
General emphatics	.74
First person pronouns	.74
Pronoun IT	.71
BE as mains verb	.71
Causative subordination	.66
Indefinite pronouns	.62
General hedges	.58
Amplifiers	.56
Sentence relatives	.55
WH questions	.52
Possibility modals	.50
Non-phrasal co-ordination	.48
WH clauses	.47
Final prepositions	.43
(Adverbs	.42)

Factor 1 (continued)	
Informational Production (–)	
Nouns	– .80
Word length	– .58
Prepositions	– .54
Type/token ratio	– .54
Attributive adjectives	– .47
(Place adverbials	– .42)
(Agentless passives	– .39)
(Past participial postnominal clauses	– .38)
Factor 2	
Narrative Discourse (+)	
Past tense verbs	.90
Third person pronouns	.73
Perfect aspect verbs	.48
Public verbs	.43
Synthetic negation	.40
Present participial clauses	.39
Non-narrative Discourse (–)	
(Present tense verbs	– .47)
(Attributive adjective	– .41)

Factor 3	
Situation-dependent reference (+)	
Time adverbials	.60
Place adverbials	.49
Adverbs	.46
Elaborated Reference (–)	
WH relative clauses on object position	– .63
Pied-piping constructions	– .61
WH relative clauses on subject position	– .45
Phrasal co-ordination	– .36
Nominalizations	– .36
Factor 4	
Overt Expression of Argumentation	
Infinitives	.76
Predictive modals	.54
Suasive verbs	.49
Conditional subordination	.47
Necessity modals	.46
Split auxiliaries	.44
(Possibility modals	.37)

Factor 5	
Abstract Style	
Conjuncts	.48
Agentless passives	.43
Past participial adverbial clauses	.42
BY passives	.41
Past participial postnominal clauses	.40
Other adverbial subordinators	.39
Factor 6	
On-line Informational Elaboration Marking Stance	
THAT clauses as verb complements	.56
Demonstratives	.55
THAT relative clauses on object positions	.46
THAT clauses as adjectives complements	.36
(Final prepositions	.34)
(Existential THERE	.32)
(Demonstrative pronouns	.31)
(WH relative clauses on object position	.30)

Table 1: Biber's factors (Biber, 1988)

Another aspect is that "it is often difficult, or indeed misleading, to concentrate on specific, isolated markers without taking into account systematic variations which involve the co-occurrence of sets of markers" (Brown and Fraser 1979:38) and, for this reason, we are following a methodology that can group the linguistic features which occur together.

We followed a methodology that emphasises the co-occurrence of linguistic features (see Biber, 1988). The main difference between our research and Biber's is that we are dealing with a very specialised corpus whereas he used texts from two large corpora containing several genres. We can say that the amount of linguistic variation in our corpus is more restricted than in his corpus and, in a certain way, our results are a simplification of what he got, but our analysis managed to capture a more fine-grained distinction.

The rest of this paper is organised in the following way. In section 2 we introduce the aspects of the methodology we are following. In section 3 we present the specific aspects of our research, in particular our corpus, and the results we obtained from the analysis. In section 4 we compare our results with other studies and, in section 5, we conclude by presenting possible ways for continuing the research.

2. Methodology

In our investigation we are following the approach described in (Biber, 1988) which tries to obtain dimensions of linguistic variation based on the grouping of linguistic features according to their correlation. The main difference between our research and Biber's is that we are dealing with a very specialised corpus whereas he used texts from two large corpora (the LOB corpus, for written texts, and the London-Lund corpus, for spoken texts) in order to have a selection of texts covering a vast range of situational uses so that he could test hypotheses about the linguistic differences between the written and spoken modes. In total he collected 481 texts comprising 23 genres.

The approach can be summarised by the following sequence of steps:

1. a large number of linguistic features are counted for each text;
2. the counts are normalised to a text length of 1,000 words (in order to make comparisons between texts possible) and also standardised to a mean of 0 and standard deviation of 1 (in order to make comparisons between features possible);
3. a statistical technique knowing as factor analysis is performed on the correlation matrix of those linguistic features aiming to group the features into a small number of sets (*factors* or dimensions[1]);
4. the obtained factors are interpreted based on previous knowledge of the linguistic variables to see if their co-occurrence makes conceptual sense;
5. scores on the factors are produced for each text so that comparisons between texts are possible.

Factor analysis is a statistical technique normally used to uncover the underlying relations/structures behind a large set of variables. It allows for the reduction of the dimensionality of the variable's space into a small set of *factors* by grouping those variables that correlate. Each factor is formed by a linear combination of the original variables.

Biber (1988) obtained six factors (presented in table 1). Positive and negative sides (marked by 'plus' and 'minus' signs on the factor labels) characterise opposing groups of co-occurrence. Labels (in italics) represent the interpretation Biber ascribed to each factor.

[1] 'Factor' and 'dimension' will be used interchangeably in this paper.

3. Our corpus and results

The corpus, tagging procedure and search algorithms

We are dealing with a corpus of 342 different texts obtained from a compendium of patient pharmaceutical leaflets (ABPI, 1996). The leaflets' main purpose is to instruct the patients on how to use the medicine, present its composition, provide possible reasons for why they are taking it, alert them to possible side-effects and to actions they should follow in case any side-effect occurs, and so forth.

The counting process was done in two stages. First, we used Brill's part-of-speech tagger (Brill, 1994), with post-correction done by substituting the words that were consistently tagged wrongly[2]. Second, we used programs written in AWK (Aho *et al.*, 1987) to count the specific patterns we were looking for — we implemented programs to count all 67 linguistic features Biber used in his analysis[3] (see appendix 2 of (Biber, 1988) for the list of features, and algorithms).

Although Biber's features comprise a reasonable size list, we added a few extra features for experimentation (the first four features), but also for comparison with other results (the last two):

- *sentence length*, measured in words by the total number of words of a text divided by the number of lines;
- *paragraph length*, measured in two different ways: (1) it was measured in words by the total number of words divided by the total

[2] We collected a list of more than 150 words, most of them nouns that were marked as a type of verb; examples include everyday words like 'pregnant' and 'either', but mainly more technical terms like 'fluoxetine' and 'hydroxypropil'.

[3] Our implementation tried to follow Biber's algorithms as closely as possible but some tweaking was necessary for some features. Those features whose counts were not reliable (and which we could not find any way to improve) were excluded from the analysis (for instance, 'past tense' and 'pro-verb do' were excluded for this reason). Other variables were excluded for reasons related to the factor analysis technique.

Orthogonal Rotated (Varimax) Factor Pattern			
Linguistic features	Factor 1	Factor 2	Factor 3
1st and 2nd pronouns	.73	.39	.
Core verbs	.65	.48	.
Conditionals	.59	.	.
Imperatives	.40	.41	.
Nominalizations	-.57	.	.
Agentless passives	-.65	.	.
Prepositions	-.71	.	.
IT pronouns[4]	.	.59	.
Private verbs	.	.55	
Infinitives	.	.50	.
Generic nouns	.	.46	.
Attributive adjectives	-.32	-.46	.
Word length	-.31	-.55	.
Nouns	.	-.75	.
Sentence length	.	.	.96
Paragraph length (measure in words)	.	.	.82
Commas	.	-.36	.72
NOTE: Values less than 30 have been printed as '.'.			

Table 2: Our factor analysis solution.

number of paragraphs (measured by number of paragraph breaks); (2) it was measured in sentences (see above), by the total number of sentences divided by the total number of paragraphs;

- *commas*, measured in relation to 1,000 words;
- verbs in the *imperative* mood;
- a list of *core verbs* taken from (Sigley, 1997): 'begin', 'come', 'feel', 'find', 'get', 'give', 'go', 'keep', 'know', 'let', 'look', 'make', 'put', 'see', 'start', 'take', 'think', 'use', 'want', and their derivatives (e.g., for 'begin', they are 'began', 'beginning', 'begins', 'begun');
- a list of *generic nouns* taken from (Sigley, 1997): 'anyone', 'anything', 'bit', 'everything', 'group', 'kind', 'lot', 'ones', 'others', 'people', 'place(s)', 'somebody', 'someone', 'something', 'somewhere', 'sort', 'thing(s)', 'time(s)', 'way(s)'.

Results of our factor analysis and its interpretation

Table 2 presents the results we obtained from our factor analysis. The first aspect to mention is the reduction on the number of variables retained in the analysis. The second aspect is that, in this solution, the factors are not correlated with each other (they are orthogonal).

Table 2 should be interpreted in the following way. Only values above .30 are presented as significant (Gorsuch, 1974:186). The value in each cell presents the correlation of the linguistic feature with the factor (called the *loading* on the factor). In each column, the positive and negative signs distinguish the variables which co-occur together — in *factor 1* (positive side) those variables are *1st&2nd Pro*, *core verbs*, *conditionals* and *imperatives*.

Interpreting the factors

Starting with *factor 1*, the positive loadings mark involvement with the reader (*1st&2nd pros*[5]),

[4] In our counts of *IT pronouns* we excluded those forms where 'it' has no referent (for instance, "it is important that …").

[5] *1st person pronouns* (PRO1), in our corpus, are used with the same discourse function of *2nd person pronouns* (PRO2) — they are written, always in questions, as if the reader himself were producing the sentence (e.g., "Should *I* be receiving this medicine?", "What do *my* tablets contain?").

and an explicit connection between possibility and action (*conditionals* and *imperatives*), characterising a direct way of giving instructions which focus on the interaction with the reader ("involved discourse" is the term used by (Chafe, 1982) for this kind of interactive texts). On the other side (negative loadings), there is a tendency towards abstraction (*agentless passives* and *nominalizations*[6]) and integration of information and qualification (by the use of *nominalizations* together with *prepositions* and *attributive adjectives*). This factor can be labelled as 'involvement *versus* abstraction'. As an example, look at the texts presented in figure 1: for the company *Organon*, Livial is the representative of its involved pole (scoring +0.49), whereas Normegon is a representative of its abstracted pole (scoring –1.89) — the range of scores on *factor 1* for the texts from the section on 'how a patient should take his medicine' goes from –2.85 to +2.24 .

On *factor 2*, the negative loadings mark a more specific, qualified type of reference (*nouns*, *longer words*, and *attributive adjectives* — possibly forming lists of referents by using *commas*) defining a more nominal style whereas, on the positive side, the loadings mark a less explicit reference (*IT pronouns* and, to a lesser extent, *1st&2nd pros*), less specific reference (*generic nouns*) and, in addition, a more verbal style (*imperatives*, *core* and *private verbs*, and *infinitives*). This factor can be labelled as 'nominal style, and explicit referencing *versus* verbal style, and pronominalised referencing'. As an example, look at the texts in figure 2: for the *Glaxo* company, Efcortelan can be considered an unmarked text with respect to *factor 2* (scoring –0.02), whereas Ceporex is at one extreme of the scale (scoring +2.15) — the range on *factor 2* goes from –2.42 to +2.15 for the texts from the same section mentioned for *factor 1*.

Factor 3 seems too specific and, for lack of space, we will not discuss it here.

4. Comparing analyses

Our comparison will be based on Biber's results presented in table 1 and ours presented in table 2. The first point to notice is that only a small number of linguistic features are prominent in our analysis and this can be attributed to the corpus we are studying. The dropping of features was done based on statistical grounds. Biber did not drop features from his analysis, although he should have done (Lee, 1999). Nonetheless, his analysis would still end up with much more features than ours (this can be attributed to difference between our corpora).

An examination of both tables 1 and 2 will show that with the exception of three features (*conditionals*, *infinitives*, and *nominalizations*) all of our features occur on Biber's factor 1. *Conditionals* and *infinitives* occurred on Biber's factor 4 "overt expression of argumentation", and *nominalizations* occurred on Biber's factor 3 (negative side) "elaborated reference". The interesting point is that our analysis obtained a finer distinction than that of Biber's factor 1, in the sense that those features are here split into two factors. With the exclusion of a lot of features which were not playing a reliable role in our analysis, the dimensions are even clearer: our *factor 1* opposes abstraction (negative side) to involvement/directness (positive side), and our *factor 2* opposes full reference (negative side) to pronominalised reference (positive side), in addition to opposing nominal to verbal style. In comparison, just the positive side of Biber's first factor is characterised by him as "verbal,

---

[6] *Agentless passives* create a sense of impersonality since there is no specified agent for the action. *Nominalizations*, in addition, can push this notion further by leaving out other parts of a proposition. Compare 1 to 4: 1) *the reviewers criticised his play in a hostile manner;* 2) *the reviewers' criticism of his play*; 3) *the reviewers' criticism*; and 4) *the criticism* (from Quirk *et al.* (1985:1289)).

interactional, affective, fragmented, reduced in form, and generalized in content" (Biber, 1988:105).

Another study applying a similar methodology to Biber's was conducted by Sigley (1997) using 3 corpora of New Zealand English with sampling characteristics similar to Biber's study — i.e., texts from several genres were used, including written and spoken texts — but using a restricted set of linguistic features. Another difference was that he used principal component analysis (instead of factor analysis). He obtained three principal components (PC)[7] but interpreted just the first two: PC 1 was considered as "a combination of [factors] 1 and 3 from Biber's (1988) analysis, and PC 2 similar to [factor] 5" (Sigley, 1997:218). He characterised his first component as measuring a notion of formality.

In comparing our result to Sigley's, we can reach the same conclusion presented above: our analysis obtained a more specialised division of functionality when looking at the factors separately. However, if we consider that some variables have loadings in both factors, implying that there is a relationship between our first two factors and they could be analysed in a higher level, this higher-level interpretation (combining the first two factors) is akin to Sigley's PC 1 interpretation.

5. Conclusions and future work

We presented an analysis of linguistic variation that uses a factor analytic technique to group the variables responsible for the bulk of the variation found in our corpus. Two main factors were found and analysed in more detail; they also were compared with those from two other analyses (Biber's and Sigley's analyses).

As far as we know, this is the first study to use Biber's methodology in a restricted corpus which started with counts for all linguistic features used by Biber (1988). The division of (mainly) one of Biber's factors into two in our analysis may be a sign that other types of specialisation could be found if other genres were factor analysed separately, and thus a better mapping of factors/linguistic features can be obtained, possibly with the creation of a hierarchy of factors (as those found in the psychology field (Gorsuch, 1974), and perhaps this is the way forward since more general factors are doubtful to appear[8]. Furthermore, it also shows that if we tried to use Biber's results to score our texts, we could end up with a blurred classification since a lot of noise would be introduced by the use of linguistic features which are not important to our specific corpus.

There are several avenues for continuing this work. Perhaps the most interesting is to seek for the reasons why a certain leaflet was written in a given style than another. For example, leaflets of solution for injection in the section on 'how to take the medicine' tend to be written with an abstract style (cf. our factor 1). A possible explanation is that for those medicines the patient is not the direct responsible for the action — the injection will be applied on him (possibly by a doctor or nurse). Nonetheless, the same kind of explanation cannot be found for tablets for instance. Other variables, like medicine group or its indication, should be used to try to get a better model.

Acknowledgements

I would like to thank David Lee, Roger Evans, Donia Scott and Richard Power for comments on earlier versions of this paper.

This work was financially supported by a studentship I received from University of Brighton.

[7] 'Component' is the correspondent to 'factor' in factor analysis.

[8] And even some of Biber's factors may not be as general as he claims — David Lee (1999) could not replicate all of them using a corpus of contemporary English with more than 650 texts including several written and spoken genres.

References

ABPI (1996) *ABPI Compendium of patient information leaflets*. Datapharm Publications.

Aho, A.V., Kernighan, B.W., and Weinberger, P.J. (1987) The awk programming language. Addison-Wesley.

Biber, D. (1988) *Variation across speech and writing*. Cambridge University Press.

Brill, E. (1994) Some advances in rule-based part of speech tagging. In *Proceedings of AAAI'94*, Seattle.

Brown, P. and Fraser, C. (1979) Speech as a marker of situation. In (eds.) Shcerer, K.R., and Giles, H., *Social markers in speech*. Cambridge University Press, pp. 33-62.

Chafe, W. L. (1982) Integration and involvement in speaking, writing, and oral literature. In Tannen, D. (ed.) *Spoken and written language: exploring orality and literacy*. Ablex. pp. 35–54.

Gorsuch, R. L. (1974) *Factor analysis*, 1st edition. Saunders.

Hair, J.F. Jr., Anderson, R.E., Tatham, R.L., Black, W.C. (1998) *Multivariate data analysis*. 5th edition. Prentice Hall.

Lee, D. (1999) *Modelling variation in spoken and written language: the multi-dimensional approach revisited*. Ph.D. Thesis. Lancaster University.

Paiva, D.S. (1999). Investigating NLG Architectures: Taking Style into Consideration. *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics* (EACL'99), student session, Bergen, Norway, pp. 237–240.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985) *A comprehensive grammar of the English language*. Longman.

Sigley, R. (1997) Text categories and where you can stick them: a crude formality index. *International Journal of Corpus Linguistics*, Vol. 2(2), pp.199-237.