Textual typology and interactions between axes of variation

Sylvain Loiseau

Textual typology and interactions between axes of variation

Sylvain Loiseau

1 Introduction

This article aims at bringing some aspects of variationist frameworks into text typology and corpus-based analyses of variation. Textual typology is an approach for describing language variation specific to corpus linguistics. The question of text typology is rooted in an old philologic and literary tradition; however, in the methodological context of corpus linguistics, it has become a method for studying the general linguistic question of language variation. It may be described with four properties. First, it considers the text as a key unit, and it accounts for regularities and correlations at the level of the text. Second, it uses statistical analyses in order to make texts comparable and summarise large amounts of data. Third, it implies using corpus methodologies in order to build and search large corpora. Last of all, it uses well-established notions as categories of variation – genres, registers, or text types – giving them a slightly different meaning.

Variationist linguistics has proposed a distinction between several axes of variation. For instance, Flydal (1952), Weinreich (1954) and Coseriu (2001) – see also Völker 2009 for a recent presentation – have distinguished between up to four axes of variation: variation across space, time, socio-cultural background and situational position: "[…] a natural language is not a homogeneous system: it is a collection of different systems, which are more or less overlapping […]. In a language, there are well known differences according to space (diatopic), according to sociological and cultural groups of the community (diastratic differences), and differences according to expressivity, following the situation type and the way of speaking, differences that I call diaphasic" (Coseriu 2001: 112)[1]. Other types of variation have been analyzed, such as the "conceptional" variation (Koch and Oestereicher 2001), which accounts for the degree of spontaneity/personal implication of the speaker. Finally, variationist linguistics includes the concept of genre: "in letters, commercial negotiations,

1. "[…] une langue historique n'est pas un système homogène: c'est une collection de systèmes différents qui coïncident en partie et en partie se distinguent les uns des autres […]. Dans une langue historique, il y a les différences bien connues dans l'espaces, ou diatopiques, et aussi des différences entre les couches socio-culturelles de la communauté (différence diastratique) et des différences entre les modalités expressives déterminées par les types de situations de l'activité de parler, différences que j'appelle diaphasique."

poetry or scientific text, speakers are recycling pieces of previous utterances belonging to the same textual genre and use a large inventory of prefabricated linguistic materials." (Glessgen 2007 : 104)[2]. In sum, in this framework, "every utterance is simultaneously localized in three dimensions: variationist, conceptional, and textual" (Glessgen 2007: 106)[3].

In this article I will argue that text typology based on large corpora and statistical methods may benefit from the notion of the plurality of axes of variation as described in variationist frameworks. In the last two decades numerous statistical text classification experiments have been proposed, starting mainly with Biber (1988). These experiments have shown that there is variation across several levels of analysis: lexicon, morphosyntax, but also morphological (Baayen 1994) or prosodic features (Obin et al. 2008). They have also shown that variation occurs across several descriptive categories: texts have been shown to vary according to genre, discourse, domain, author style, but also according to socio-geographic variables (van Keune and Baayen 2006), modality, i.e. speech/writing (Biber 1988, Plag et al. 1999), "text type" (Biber 1988, Baayen 1994), etc. Morphosyntactic tagsets of different granularity have been used and a high number of statistical methods have been tested and evaluated.

The limits of automatic text classification for textual typology, however, are well known. Little consensus has emerged as to how to define and stabilize these descriptive categories (text types or genres). Many general, common sense, broad categorisations may be illustrated with statistics based on large corpora. This is especially true if the corpus is organized in several very different groups of texts. For instance, Obin et al. (2008) succeeded in automatically classifying texts into five different "discourses". But these discourses were very different: "radio news", "task map", "political discourse", "life story" and "radio interviews". It mixes oral and written texts (or oralized texts, cf. Koch and Oesterreicher 2001), different degrees of spontaneity, genre, theme… The result of the experiments does support the hypothesis that prosodic features are discriminatory (this was the aim of the paper). However it does not entail any better understanding of linguistic variation. In such a classification, one may argue that we find the categories we have put in the corpus.

More generally, the fact that the statistical classification is successful does not entail that the typology is scientifically grounded or that we gain better knowledge of the units (genres) from it. Kilgarriff (2005) showed that virtually every statistical textual classification experiment, whatever the parameters may be, shows that the distribution of features is non-random, without giving evidence that it is non-arbitrary: "the probability model, with its assumptions

2. "Pour une lettre ou une conversation d'achat comme pour une poésie ou un texte scientifique, les énonciateurs reproduisent le modèle d'autres discours semblables appartenant au même genre textuel et ils puisent dans un vaste inventaire d'élèments de langue préfabriqués."
3. "Tout énoncé s'inscrit donc parallèlement dans les trois dimensions, variationnelles, conceptionnelle et textuelle, qui sont à tout moment co-présentes."

of randomness, is inappropriate, particularly where counts are high (e.g., thousands or more)" (2005 : 268). Using a text-typology experiment, the author has shown that "given enough data, H_0 [the hypothesis that two subcorpora are distinguishable from two subcorpora which have been randomly generated on the basis of the frequencies in the joint corpus] is almost always rejected however arbitrary the data" (2005: 268). Hence, "There is no a priori reason to expect words to behave as if they had been selected at random, and indeed they do not. It is in the nature of language that any two collections of texts, covering a wide range of registers (and comprising, say, less than a thousand samples of over a thousand words each) will show such differences." (2005: 269f.). Inductive typology and typology using mainly internal properties of texts are particularly concerned by this drawback. In sum, if everything seems to vary, whatever the corpora, the linguistic features, and the statistical methods may be, are we identying and characterising a variation in a linguistic sense?

2 Hypothesis

In this paper I will explore the hypothesis that taking into account the plurality of axes of variation may be useful in textual typology. In a certain sense, the frequency of a single feature (say, the frequency of a modal verb) cannot be assigned to an axis of variation while disregarding the other axes. There is no way that we know if a feature is characteristic of, say, an author, without taking into account the properties of the genre or the discourses this author uses. An axis of variation cannot be described in isolation. The question of the relation and interaction between axes of variation is well known in variationist linguistics. For instance there is a well known relation between diastratic and diachronic variation (under some circumstances, the further you are from innovative centers, the more archaic your variety is), and between diastratic and diaphasic variation (Finegan and Biber 2001, Dufter and Stark 2002: 89, Gadet 2003). To a certain extent, classifying texts into one axis of variation – for instance, the genre – without taking into account other relevant axes of variation – such as authorship or domain, is like trying do describe diastratic variation without taking into account the age of the speakers, diatopic or diaphasic properties.

In corpora, a homogeneity regarding an axis of variation can never be "isolated" from heterogeneity through other axes. For instance, representing an idiolect in a corpus requires sampling many genres, dates, or conversational parameters. Manning (2002: 294) stresses that "there is no easy answer to the problem of getting sufficient data of just the right type: language changes across time, space, social class, method of elicitation, etc. There is no way that we can collect a huge quantity of data (or at least a collection dense in the phenomenon of current interest) unless we are temporarily prepared to ride roughshod over at least one of these dimensions of variation." This entails that

there is no "homogeneous" corpus, and, moreover, that taking into account interactions between axes of variation is required for characterising an idiolect or a genre.

3 Methodology

In order to investigate these inter-relations, I have focused on the question of which features are specific to one axis of variation, and which features are common to several axes of variation, in a corpus where several dimensions of variation are known. I have performed four independent automatic classifications, corresponding to four axes of variation on a corpus. I use a family of statistical methods, the decision trees, allowing an easy extraction the sets of discriminant morpho-syntactic features on each axis. I analyse the intersection between the sets of features. Some features are common to several sets of features and, then, shared between several axes of variation, while other are specific to one axis of variation. Each feature may be analyzed in the light of the axes it helps to predict. Can we distinguish between features specialized in one axis of variation, and features varying according to several axes of variation? Does considering features specific to one axis of variation help for characterising this axis of variation? Does analysing intersection between sets of features of each axis of variation help determine the correlation between axes of variation?

The corpus has been carefully designed in order to allow for this experiment. I used articles from the French daily newspaper *Le Monde*. Thanks to the meta-information available for each article, many dimensions can be observed:

- date: the articles available range from 1987 to 2002
- author
- genre (interview, biography, analysis, etc.)
- section (international, national, sport, opinion, enterprise, etc.)

In order to create a balanced corpus, representing various genres, authors, sections and spans of time, I selected a subset of 840 articles. This implied much pre-processing in order to deal with changes in section or genre names: a section or a genre may have different names across time due to the editorial evolution of the newspaper (see Figure 1, left). Such renaming was identified using a Hierarchical Ascendant Classification of subcorpora of articles of each section (Figure 1, right). This classification shows that some pairs of sections that are consecutive in time are also very close regarding their lexical content.

Moreover, in order to observe relationships between features and axes of variation, I have designed a corpus with as little attraction between categories as possible. For instance, I have selected articles by authors writing from 1987 to 2002, articles belonging to genres that are not specialized in only one section, etc. Eliminating as far as possible correlation between categories leads to

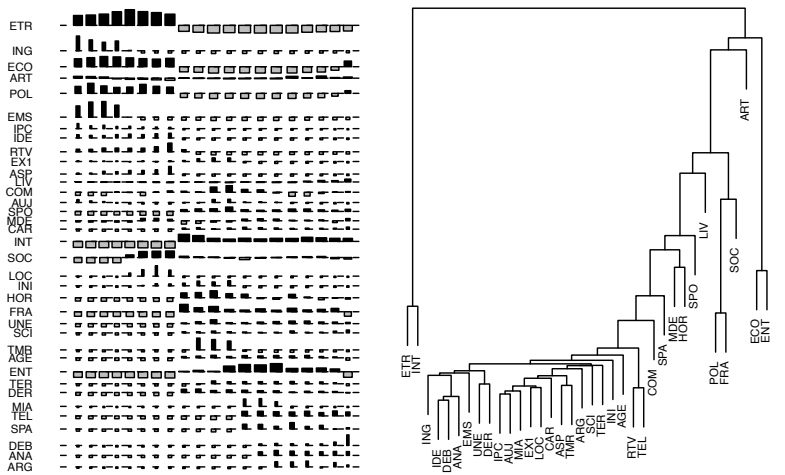

Figure 1: Left: Association plot of the sections (only sections with more than 1000 articles are kept) over the 16 years of the corpus. The section «ETR» (étranger: foreign affairs; first line) is associated with years until 1994; then the section named «INT» (international, international, middle) is associated with the following years.– *Right:* Dendrogram of a clustering of the subcorpora representing these sections (using lemma as features). Consecutive sections in the association plot are also grouped in the dendogram (ETR and INT at the far left of the plot, ECO (economy) and ENT (entreprise) at the far right, for instance).

selecting a very restricted subset of articles (840 out of 950 000 articles) containing only some authors, genres and sections of the newspaper. The 16 years were regrouped in three groups of years in order to increase the number of articles in each diachronic category. Eventually, the following categories may be observed in the corpus:

- three periods: (from 1987 to 1991 (147 articles), from 1992 to 1996 (247 articles), from 1997 to 2002 (446 articles).
- 17 authors, ranging from 4 to 120 articles.
- 4 genres (interview, biography, portrait, obituary), ranging from 56 to 350 articles. There is a strong bias due to the fact that only genres closely related to portraying poeple are represented. This is due to the fact that these genres are the only ones spreading across sections and authors.
- 6 sections (ART (art), ECO-ENT (economy and enterprise), HOR (opinion), INT-ETR (international news), POL-FRA (national news), SPO (sport), ranging from 13 (SPO) to 280 articles (INT-ETR).

Of course variables (year, genre, section, author) are not independent from each other: the chi square test reveals some attraction between every pair of tasks. The mosaic plot (Figure 2) shows associations between authors and sections on the one hand, and authors and periods, on the other hand. Nevertheless, selected authors are not completely specialized in one section or one period, and so it was the optimal "cross-balanced" corpus than could be extracted from the whole corpus.

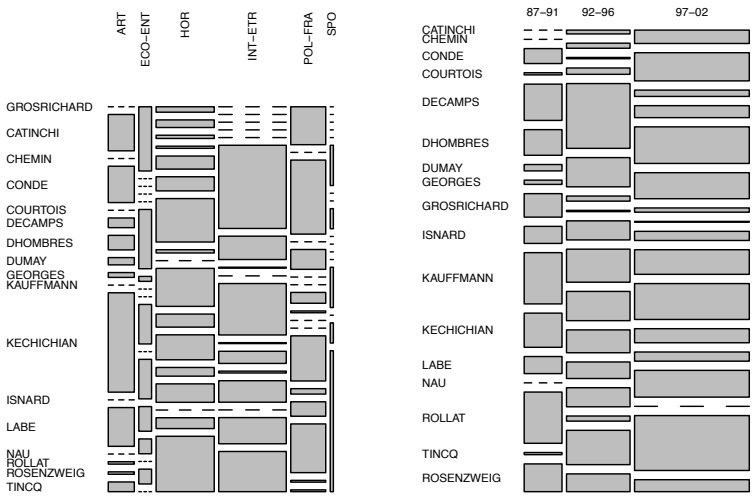

Figure 2: Mosaic plot of the associations between author and sector (left) and author and year (right)

Our corpus has been analysed using the Syntex parser (Bourigault et al. 2005). The morphosyntactic tagset of this analyser contains only 93 tags; these coarse-grained categories allow for strong robustness.

4 Results

The classification tree algorithm (Ripley 1996) is used for extracting the features making these axes predictable for each of the four axes of variation. Figure 3 shows the classification trees for two tasks: genres and sections.

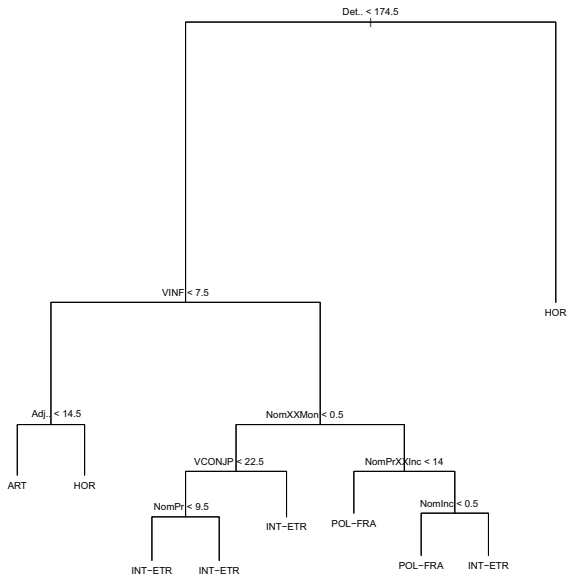

Figure 3: Classification trees for classifying articles into sections

Intersection between sets of features may be summarised as follows:

a. 73 features are never used;
b. 13 features are used for one task only (specific):
   – *Author:* CCoordAdj, CCoordPrepDe, PpaMS, Prep, PrepDet, Typo
   – *Genre:* Det, ProRel
   – *Section:* Adj.., Det.., NomInc, VINF, NomXXMon
   – *Year:* none
c. Seven features are shared between two tasks:
   – *Genre and author:* CSub, NomPrXXPrenom
   – *Genre and year:* Elim, NomFS, NomXXDate
   – *Genre and section:* NomPrXXInc
   – *Author and section:* NomPr
d. One feature, VCONJP, is shared by three tasks (genre, section, year).

Without much surprise, all features related to proper nouns (begining with "Nom-", noun) are used mainly for discrimination between section and genre, i.e. the most thematic axes of variation, and less frequently for discriminating author and year. Section and genre use different proportions of nouns. Eight features are necessary in order to distinguish between authors, while only four features are necessary to distinguish between years. Nine features are necessary in order to distinguish between genres, few of them being specific to genre: this axis of variation is mainly associated with features shared with other axes of

variation (this does not necessarily entail that it is strongly correlated to other axes of variation). Author variation, on the contrary, is the most specific axis of variation.

5 Conclusion

This experiment supports the hypothesis that some features are discriminatory for several axes of variation. Discriminatory features are not specific or "specialized" into one axis of variation; on the contrary, the same restricted set of features are used by many axes of variation. This implies that one cannot use automatic text classification for text typology, where the classes are different values along one axis of variation, without taking into account the interaction of axes of variation: there is no set of features that is discriminatory for one axis without being influenced by the others.

Further studies for a better understanding of the interaction between axes, as well as for a better understanding of the way of using these interactions for textual typology, may benefit from some machine learning algorithm supporting the classification into several sets of classes simultaneously, such as the machine learning algorithm family called "multi task learning". More generally, the analysis of these interactions seems to be a fruitful avenue for future research.

Acknowledgments. This work has been supported by TGE Adonis, CNRS.

References

Baayen, R.H.
1994 "Derivational Productivity and Text Typology", in: *Journal of Quantitative Linguistics*, 1; 16–34.

Biber, D.
1988 *Variation across speech and writing*. Cambridge: Cambridge University Press.
1993 "Using Register-Diversified Corpora for General Language Studies", in: *Computational Linguistics*, 19/3; 219–241.
1990 "Methodological issues regarding corpus-based analyses of linguistic variation", in: *Literary and Linguistic Computing*, 5/4; 257–270.

Bourigault, D.; Fabre, C.; Frérot, C.; Jacques, M.-P.; Ozdowska, S.
2005 "Syntex, analyseur syntaxique e corpus". In: *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan.

Brunet, É.
2006 "Le corpus comme une boule", in: *Proceedings Albi 2006 conference*. [Electronic source: `www.revue-texto.net/Parutions/Livres-E/Albi-2006/Brunet.pdf`]

Coseriu, E.
2001 *L'homme et son langage*. Paris: Peters.

Dufter, A.; Stark, E.
2002 "La variété des variétés: combien de dimensions pour la description?", in: *Romanistisches Jahrbuch*, 53; 81–108.

Finegan, E.; Biber, D.
2001 "Register Variation and Social Dialect Variation: the Register axiom." In: Eckert, P.; Rickford, J. R. (eds.), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 235–267.

Flydal, L.
1952 "Remarques sur certains rapports entre le style et l'état de langue", in: *Norsk Tidsskrift for Sprogvidenskap*, 16; 241–258.

Gadet, F.
2003 "La signification sociale de la variation", in: *Romanistisches Jahrbuch*, 54; 98–114.

Glessgen, M.-D.
2007 *Linguistique romane. Domaines et méthodes en linguistique française et romane*. Paris: Armand Colin.

Kilgarriff, A.
2005 "Language is never ever ever random", in: *Corpus Linguistics and Linguistic Theory*, 1/2; 263–276.

Koch, P.; Oesterreicher, W.
2001 "Langage parlé et langage écrit." In: Holtus, G.; Metzeltin, M.; Schmitt, C. (eds.), *Lexikon der Romanistischen Linguistik*. Tübingen: Max Niemeyer Verlag, 584-627.

Loiseau, S.
2008 "Corpus, quantification et typologie textuelle", in: *Syntaxe et sémantique*, 9; 73–85.
Manning, C.D.
2002 "Probabilistic syntax." In: Bod, R.; Hay, J.; Jannedy, S. (eds.), *Probabilistic Linguistics*. Cambridge: The MIT Press, 289–341.
Obin, N.; Lacheret, A.; Veaux, C.; Rodet, X.; Simon, A.-C.
2008 "A Method for Automatic and Dynamic Estimation of Discourse Genre Typology with Prosodic Features." In: *Interspeech 2008*, Brisbane, Australia.
Plag, I.; Dalton-Puffer, C.; Baayen, R.H.
1999 "Morphological productivity across speech and writing", in: *English Language and Linguistics*, 3; 209–228.
Ripley, B.
1996 *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.
van Keune, H.; Baayen, R.H.
2006 "Socio-geographic variation in morphological productivity in spoken Dutch: a comparison of statistical techniques", in: *Actes des 8es journées d'analyse des données textuelles (JADT 2006)*, 571–581.
Völker, H.
2009 "La linguistique variationnelle et la perspective intralinguistique", in: *Revue de linguistique romane*, 73/289; 27–76.
Weinreich, U.
1954 "Is a Structural Dialectology Possible?", in: *Word*, 10; 388–400.

Contents