A semi-supervised multivariate approach to the study of language variation

Sascha Diwersy, Stefan Evert and Stella Neumann

1. Introduction

The use of aggregational techniques in this paper is realized first and foremost by developing and testing a 7-step methodology that applies semi-supervised multivariate techniques to the study of linguistic variation of different sorts. This methodology is exemplified in two case studies representing different research questions with diverging requirements on the methodology. We argue that the common core methodology can be adapted to different situations with flexible parameter settings and allows researchers to investigate more complex, hidden structure beyond obvious quantitative indicators observable with the help of conventional corpus-based methods.

Multivariate techniques are used in many subfields of linguistics to study different kinds and aspects of language variation, as the papers in this volume document. While most such studies are based on some form of a statistical latent variable model, each branch of linguistics tends to develop a specific set of techniques and procedures, which are then applied uniformly in subsequent research. Despite the undeniable benefits of a standard methodology, there are also significant disadvantages: (i) researchers may be tempted to apply the established procedures and interpretations without considering their appropriateness for the specific data set and without exploring alternative approaches; (ii) it is difficult (or even impossible) to compare results from subfields of linguistics that use different standard procedures. Our goal in this paper is twofold: First, we suggest an extension to the established multivariate toolbox in language variation research, relying on visualization and weakly supervised learning in order to detect more subtle patterns in linguistic data than is possible with standard unsupervised techniques. Second, we aim to find out whether it is possible to develop a common core methodology that can be used for, and adapted to, a wide range of tasks and data sets. We test this approach in two case studies, making the present paper relevant to two distinct subfields of linguistics: contrastive linguistics with a variational twist in a case study of register variation in English and German originals and translations, and a variational approach to dialectology in a case study of African varieties of French. The first study – whose main focus is not on general differences between English and German – exemplifies an approach that is markedly different from the typologically-oriented papers in this volume. It is more closely related to the work of Biber (in particular Biber 1995) with which it shares the strong focus on text features.

Multivariate methods are particularly useful for inductive, corpus-driven research, where they help to discover 'hidden' structure in linguistic data. Typically, studies in this area rely on unsupervised statistical techniques, in particular a range of closely related latent variable models including factor analysis (FA), principal component analysis (PCA), correspondence analysis, and k-means or Gaussian mixtures clustering. Such unsupervised approaches are based on similarities between texts or other linguistic samples, computed from a set of quantitative features determined by the researcher. As a result, these methods face two major problems: Firstly, they focus on the major dimensions of language variation and may not detect more subtle patterns that are of particular interest to linguists. Secondly, the results of an analysis depend strongly on the choice of quantitative features included in the similarity computation, thus introducing a risk of circularity. This is especially problematic if features are derived from theoretical assumptions about patterns of interest (subverting the inductive aims of the research) or if a researcher attempts to select features that steer the analysis towards an expected result (fearing that it might otherwise be hidden by more prominent variational patterns).

In the present paper, we explore an approach to the multivariate analysis of language variation that aims to circumvent these two problems. While it is still based on latent-variable techniques (both FA and PCA), our approach puts strong emphasis on visualization and semi-supervised exploration of the feature space. In this exploratory analysis, we use a controlled amount of theory-neutral information

(e.g., in the first case study, whether a text is an original or a translation) in order to highlight fine distinctions hidden in the variational patterns (thus addressing the first problem) and to allow a data-driven selection and weighting of features (thus reducing the risk of circularity).

Since our aim is to develop a common methodological core for variational studies in all subfields of linguistics, which is at the same time still flexible enough to accommodate the specific requirements of different data sets and research questions, we apply the suggested procedure to both case studies with a minimal amount of necessary adaptations in the parameter settings. This also provides a more comprehensive view of the usefulness and potential pitfalls of our methodology than a single illustrative example.

The remainder of the paper is organized as follows. Firstly, we introduce the two case studies used to illustrate and assess the multivariate analysis (Section 2). Section 3 then describes and motivates our new methodological core in the form of a 7-step procedure. The following sections are concerned in turn with the application and adaptation of the common methodology to case study 1 (section 4) and case study 2 (section 5), as well as with the linguistic discussion of the results. Finally, section 6 summarizes the findings of the paper and gives an outlook on future work.

2. Introduction to the two case studies

This paper reports two test cases to which the same methodological procedure was applied. Their general characteristics and theoretical background are briefly introduced here. A more detailed discussion of the respective multivariate analyses will follow in sections 4 and 5.

The first case study is concerned with the identification of intra- and interlingual variation between registers in the language pair English and German. Crucially, it also concerns variation between originals and translations, as translations are assumed to be a potential gateway of language contact and thus ultimately language change (see Neumann 2011b). These aspects have been examined in a previous study drawing on frequencies of individual features rather than multivariate correlation patterns (Neumann 2008). Neumann (2008) interpreted combinations of features qualitatively and made use of statistical methods only for the purpose of significance testing.

The corpus used for this purpose is the CroCo Corpus (Hansen-Schirra, Neumann and Steiner to appear), a resource consisting of 462 texts[1] from eight different registers (political essays, fictional texts, instruction manuals, popular-scientific writings, letters to shareholders, prepared speeches, tourism brochures, and webpages) in English and German. For each original text the corpus also includes a (published) translation in the respective other language. The total size of the corpus is 1,181,435 word tokens including punctuation marks. In addition to this core corpus, two small reference corpora are part of the CroCo Corpus consisting of 2,000 word samples from 17 different registers in both languages (not used in our case study). The corpus contains meta-information for each text, as well as morphological, part-of-speech and phrase structure annotation enriched with information on syntactic functions. Furthermore, clauses and sentences are segmented. Each translation and the corresponding original are also aligned at word, phrase, clause and sentence level, but this information is not used for the purposes of the present study.

The 29 linguistic indicators used in case study 1 were derived in a theory-driven approach (Neumann 2008) drawing on systemic functional register theory (Halliday and Hasan 1989; Martin 1992; Matthiessen 1993 etc.). Feature derivation was thus aimed at providing as broad as possible an overview of indicators for register distinction, theoretical assumptions on translation status or the distinction between the two languages under investigation were not taken into consideration for feature derivation. Note, however, that contrastive differences between functionally comparable register indicators had to be accounted for in the feature selection (see section 4.3). The 29 features comprise lexico-grammatical features as diverse as number of nouns per all tokens, number of imperatives per all sentences, number of subjects in sentence-initial, i.e. thematic position per all themes, contractions per all tokens, the type-token ratio for lexical words, etc. Frequency information for each text was

obtained from complex corpus queries based on the multi-layer annotation of the CroCo corpus and was adjusted to yield the relative values mentioned above.

The fact that this paper uses data that have already been interpreted qualitatively in a previous study offers the advantage of allowing a comparison between the linguistic interpretation of the data set in Neumann (2008) and the results of our multivariate analysis.

The second case study focuses on regional variation of the French language in post-colonial African countries. This research paradigm – which has been pursued since the mid-1980s by French scholars such as Suzanne Lafage and Ambroise Queffélec – over the years has led to many publications describing the socio-linguistic and linguistic aspects of the evolution of French, including a wide range of detailed inventories of lexical particularities in different African countries.[2] The compilation of these inventories was carried out according to rather conventional lexicological and lexicographic methods based on corpora collected for the purpose. At the same time, a large-scale investigation using methods of contemporary corpus linguistics still remains a desideratum, as stated by Stein (2003:14–15).

A corpus-driven approach to the study of post-colonial varieties of French (as well as other languages such as English, Spanish or Portuguese) should address at least the following issues:

1. To what extent is there evidence of an overall polycentric evolution, such that the emergence of endogenetic, national varieties can be established?
2. Which regions/countries stand out by showing a more innovative or conservative variational trend?
3. Which type of linguistic patterns, i.e. collocations, colligations and semantic associations in terms of Hoey's (2005) Lexical Priming Theory, are most likely to vary in a given region or country?[3]
4. Which items or classes of items differ most in their functional values as determined by their distributionally constrained combinatorial profiles[4]?

It is clear that corpus-driven work on these issues necessitates the use of rather sophisticated statistical techniques. The aim of the present paper is to contribute to the development of adequate methodological strategies based on multivariate data exploration.

The samples used in the second case study are extracted from the CPFC (*Corpus numérisé de la Presse francophone*), a corpus archive of francophone newspapers being compiled since 2010 at the Center of Interdisciplinary Research on France and the Francophone World (CIFRA) at the University of Cologne. All documents in the archive are part-of-speech tagged, lemmatized and dependency-parsed[5]. The CPFC comprises in its present state (as of February 2011) the sub-corpora organized by regional parameters listed in Table 1.

Table 1. Composition of the CPFC with respect to different countries.

Sub-corpus (country)	**Number of texts**	**Number of word tokens**
France	619,000	309,500,000
Algeria	75,200	37,600,000
Belgium	49,500	24,750,000
Benin	19,200	9,600,000
Burkina Faso	16,700	8,350,000
Congo (R.C.)	17,900	8,950,000
Congo (R.D.C.)	38,400	19,200,000
Cameroon	91,200	45,600,000
Canada (Québec)	63,000	31,500,000
Ivory Coast	42,600	21,300,000

Lebanon	53,200	26,600,000
Madagascar	20,500	10,250,000
Mali	19,700	9,850,000
Morocco	68,000	34,000,000
Senegal	36,200	18,100,000
Switzerland	52,600	26,300,000
Tunisia	32,800	16,400,000
TOTAL	1,315,700	657,850,000

As compilation and processing of the overall corpus archive indicated in Table 1 are carried out asynchronously over a longer period of time, only a part of these samples currently meets the criteria required for the present case study aimed at illustrating the new methodology. In particular, requirements were publication time (data from the same (or similar) two years) and (at least) two major newspapers from each country (which should also be comparable in terms of variety). We therefore had to discard, amongst others, the samples representing Algeria (containing 3 different newspapers, but only from single and different years) and the Democratic Republic of Congo (containing 3 years of only one newspaper).

In order to base our study on quantitatively and qualitatively balanced data, we created samples for six different countries, each consisting of approximately 14,000,000 – 14,500,000 word tokens taken from two years each of two different newspapers. Every newspaper volume was sub-divided into ca. 50 equally sized units, approximately corresponding to weekly editions. Table 2 below gives an overview of the composition of these samples.

Table 2. Data samples used in the experiment.

Sample	**Newspaper volumes**[6]	**Sample abbreviations**	**# items**	**Word tokens**
Cameroon (CAM)	*Mutations* 2007-08 *Cameroun-Tribune* 2007-08	MUTA TRIB	53 + 53 53 + 53	14,200,000
France (FRA)	*Le Monde* 2007-08 *Le Figaro* 2007-08	LM LFI	53 + 53 53 + 53	14,500,000
Ivory Coast (CIV)	*Fraternité-Matin* 2007-08 *Notre Voie* 2007-08	FRAT VOIE	53 + 53 53 + 49	14,100,000
Morocco (MAR)	*Aujourd'hui-Le Maroc* 2008-09 *Le Matin du Sahara* 2008-09	AJD MAT	53 + 53 53 + 53	14,300,000
Senegal (SEN)	*Le Soleil* 2007-08 *Walfadjri* 2007-08	SOL WALFA	53 + 53 53 + 53	14,400,000
Tunisia (TUN)	*La Presse* 2008-09 *Le Temps* 2008-09	LAPRE TEMPS	53 + 53 53 + 53	14,200,000
TOTAL	12 samples		1,268	85,700,000

From the different types of syntagmatic patterns analyzed in previous work, we chose colligations of nouns, i.e. their passive and active valency[7], as a test case. Technically speaking, for every noun token in each sub-sample, the lemma and syntactic functional label was extracted and frequency counts were calculated for each distinct lemma-function pair. Quantitative information about the database obtained by this procedure is shown in Table 3. In the experiments, only colligational pairs above a specified

frequency threshold were included, resulting in roughly 8,250 – 18,500 features (see section 5 for details).

Table 3. Database of noun colligations from the six samples used in the experiment.

Sample	**Colligation types**	**Colligation tokens**
Cameroon (CAM)	4,503,334	10,326,558
France (FRA)	4,997,851	10,248,651
Ivory Coast (CIV)	3,869,833	10,470,315
Morocco (MAR)	4,334,878	11,133,241
Senegal (SEN)	4,444,827	10,617,994
Tunisia (TUN)	4,604,645	10,735,184
TOTAL	26,755,368	63,531,943

By choosing functionally disambiguated nouns as features we want to test, on the one hand, the implementation of a method extending the well-established bag-of-words paradigm. On the other hand, the selected feature combination should yield, on the level of corpus extraction, an approximation to colligation as one of the distributional ranks proposed by Hoey's Lexical Priming Theory. The strong hypothesis related to this theory would be that samples coming from one or another region could exhibit different degrees of distributional deviation with respect to different pattern ranks (lexical collocations, semantic association, textual collocation etc). The implementation of a complete sampling procedure covering the extraction of feature candidates adequate to each of these ranks, nevertheless, goes far beyond the scope of this paper. Selecting just one feature combination is certainly a rather practical solution and can only be considered as a starting point for further investigation which should take into account feature sets addressing other distributional ranks.

A comparison of the two case studies – highlighting the differences between the data sets and annotations – is shown in Table 4 below. Trivial indicators as mentioned in the table are those that directly provide an obvious separation of groups and thus do not require advanced multivariate methods.

Table. 4. Comparison of the two case studies and the data sets used.

	English & German registers (case study 1)	**Varieties of French (case study 2)**
Feature matrix	454 texts × 29 features	1,268 texts × 8,250 – 18,500 features
A priori assumptions	Strong theoretical assumptions	(Almost) none
Annotation	Multi-layer annotation	Dependency parser
Linguistic categories	Language, register, translation status	Country, newspaper
Language	English, German	French
Features	Various lexico-grammatical features	Colligations = noun lemma + dependency
Trivial indicators	Features only existing in one of the two languages, e.g. the German modal passive	(Frequencies of) country-specific lexemes

3. Motivation and method

The goal of our current research is to develop an inductive multivariate approach for corpus-driven variational studies that (i) is able to detect subtle, linguistically relevant patterns and (ii) avoids the

risks of circularity and researcher bias introduced by selecting quantitative features based on theoretical assumptions or the researcher's expectations.

Our starting point is factor analysis (FA), an unsupervised latent-variable technique that has become highly popular in research on language variation following the work of Biber (1988). Factor analysis infers a small number of latent dimensions to account for the observed variation between texts or other linguistic samples, with respect to a set of quantitative features defined by the researcher. The underlying assumption is that the latent dimensions represent the major parameters of language variation, whereas the remaining 'unexplained' variation is due to chance, thematic differences and other text-specific properties.

For instance, we applied factor analysis to the 454 German and English texts from case study 1, each described by the relative frequencies of 29 linguistic indicators (cf. section 2). Figure 1 in section 4 shows the result of a three-dimensional FA in the form of a scatterplot matrix: each point corresponds to one text sample and its position in the panels indicates the placement of the text with respect to the three latent dimensions. For example, the first latent dimension is represented by the *y*-axis (vertical) of the panels in the first row; the third latent dimension by the *x*-axis (horizontal) of the panels in the last column.

It is obvious from Figure 1 that the factor analysis captures the difference between English and German (in the second latent dimension, e.g. on the *x*-axis of the top left panel) and the main patterns of register variation (first and third latent dimensions, which are characteristic for FICTION and INSTRUCTION MANUAL, respectively). However, the subtle differences between translated and original texts remain undetected by the completely unsupervised factor analysis approach, being obscured by much larger differences between the two languages and between individual registers.

Despite its popularity and its success in many applications, factor analysis has two methodological disadvantages for our purposes: (i) the number of latent dimensions has to be determined *a priori* by the researcher, and changes in this parameter often lead to markedly different results; (ii) because factor analysis separates correlation patterns from the independent variation of individual features (their "uniqueness"), it cannot be applied if the data set contains more features than data points (as in our case study 2).

For these reasons, we use the related but simpler technique of principal component analysis (PCA). Principal component analysis is computationally stable, can be applied to high-dimensional data sets, and determines an arbitrary number of ranked latent dimensions (called principal components, PC). In addition, principal component analysis has a geometric interpretation as an orthogonal projection into a lower-dimensional subspace that preserves distances between data points as far as possible. It can also meaningfully be combined with other coordinate transformations.

In our first case study, the scatterplots for factor analysis and principal component analysis (not shown) are strikingly similar. While the principal components are not aligned with linguistic distinctions as precisely as the FA dimensions, this is unproblematic in our visualization-oriented approach: e.g. the axis separating English and German texts is clearly visible in the PCA scatterplot matrix even if it does not coincide with a single latent dimension. Therefore, the use of principal component analysis instead of factor analysis does not result in a loss of relevant information. In our second case study, the factor analysis algorithm failed to converge due to the very large number of features, leaving no alternative to principal component analysis.

As we have pointed out above, completely unsupervised techniques such as factor analysis and principal component analysis only detect major patterns of language variation and are sensitive to the choice of quantitative features. In our two case studies, the general variation between languages, registers and individual texts completely obscures subtle distinctions that are the main focus of interest: the differences between originals and translations (case study 1) and the regional variation of French across different African countries (case study 2). Therefore, we propose the following methodological approach based on visual interpretation of latent dimensions and a weakly supervised exploration of the feature space, which carefully introduces a small amount of theory-neutral knowledge.

1. An unsupervised principal component analysis (or factor analysis) produces a ranked list of latent dimensions that account for the variation between individual text samples.
2. Latent dimensions are visualized as a scatterplot matrix or with interactive three-dimensional graphics. Linguistic categories of interest (e.g. country, register, translation status) are indicated by color, plot symbol, etc. This visualization shows whether the unsupervised analysis was able to capture the relevant linguistic distinctions; it may also highlight some potential methodological problems (cf. section 5).
3. If the desired dimensions have not been identified in step 2, a minimal amount of prior knowledge has to be included. In order to maintain the inductive nature of our approach, it is important to use only theory-neutral information. In case study 1, for example, we mark texts as originals and translations, which is independent from any theoretical assumptions about characteristic properties of translated texts. In case study 2, we group together the text samples from each newspaper. However, we do not include information about the country of origin, as this would presuppose the existence of a distinct regional variety of French in each of the six countries.
4. Based on this prior knowledge, a targeted search identifies latent dimensions that capture the patterns of variation between the specified linguistic categories. Our method of choice is linear discriminant analysis (LDA), which can easily be visualized and combined with selected unsupervised principal components.
5. Especially when large numbers of quantitative features are available, there is a danger that the linear discriminants might just consist of random selections of features that happen to correlate with the specified linguistic categories in this particular data set. Therefore, the results of step 4 need to be validated by measuring the classification accuracy of the LDA model, using cross-validation techniques in order to ensure a sound separation of training and test data. The linear discriminants are only meaningful if a sufficiently high classification accuracy is achieved. Supervised machine learning with support vector machines (SVM) provides an additional quantitative evaluation, which can also identify non-linear patterns. Note that this validation step helps to avoid circularity and deductive bias in our weakly supervised approach, because latent dimensions are not selected at the researcher's whim but rather according to their proven ability to distinguish the theory-neutral categories introduced in step 3.
6. The procedure is repeated from step 2 (visualization). If the latent dimensions are not satisfactory yet, additional prior knowledge may have to be included or other refinements may be necessary (cf. sections 4 and 5). In this way, the analysis is gradually improved through multiple iterations of the procedure.
7. When suitable latent dimensions have been found, the corresponding feature weights – combined with visualization and quantitative validation of the dimensions – form the basis of a linguistic interpretation. It is needless to say that, without such interpretation, the purely computational results of our exploratory procedure would remain meaningless.

The two case studies reported in the following sections both make use of this common procedure. However, different parameter settings, modifications and refinements are required in each case. Therefore, our methodological core should not be seen as a fixed sequence of steps to be followed mindlessly, but rather as a guideline for the weakly supervised and corpus-driven multivariate analysis of language variation. Careful inspection and interpretation of the results obtained at each step is essential for a successful application of the procedure.

4. Case study 1: English and German registers in original and translation

4.1. Multivariate procedure

As introduced in section 2, this study operates on a data matrix of 454 texts × 29 features. The definition of features is based on theoretical assumptions. For example, lexico-grammatical features are

derived from the assumption that a register reflects three main strands of meaning, namely referential meaning (field of discourse), pragmatic meaning (tenor of discourse) and a specific linguistic organization (mode of discourse; cf. Halliday and Hasan 1989). The features are quantified as relative frequencies and given in the form of percentages (for a detailed description of the data collection procedure, see Neumann 2008). Even though most features are percentages and thus on the same scale of measurement, their value ranges are strikingly different. Since some multivariate analysis and visualization techniques are sensitive to scaling, we transformed all features into standardized z-scores.

There are strong correlations between some of the variables, which have to be taken into consideration in the discussion of the multivariate analysis (cf. section 4.2). The following examples are also interesting from a linguistic point of view:

- Nouns, attributive adjectives, nominalizations and lexical density are negatively correlated with subordination and pronouns. This appears to reflect the distinction between texts containing spoken language and those completely situated within the written mode.
- It is by no means surprising that contractions and colloquialisms are correlated since they can be assumed to appear in combination in casual interactions. There is a somewhat weaker correlation of these two features with finites. This further corroborates the indication of casual conversation as this can be assumed to take place in the spoken mode which is typically described as consisting of more (and shorter) clauses (see, for instance, Halliday 2001).
- Finally, imperatives and verbs as theme are correlated. Obviously, the verb is placed in the sentence-initial and thus thematic position in the imperative mode, displacing the subject. Hence, subject in thematic position is correlated negatively with imperative mode and verbs as theme.

In step 1 of the procedure described in section 3, we carry out a three-dimensional factor analysis with varimax rotation and regression scores as factor coordinates in order to infer latent properties from the feature correlations. In step 2, a scatterplot matrix of the regression scores (see Figure 1) shows that factor analysis separates German and English texts reasonably well (second latent dimension, marked by rectangles in Figure 1) and that two registers (FICTION and INSTRUCTION MANUAL) clearly contrast with the other registers which, in turn, do not display any clear separation (cf. circles in Figure 1).

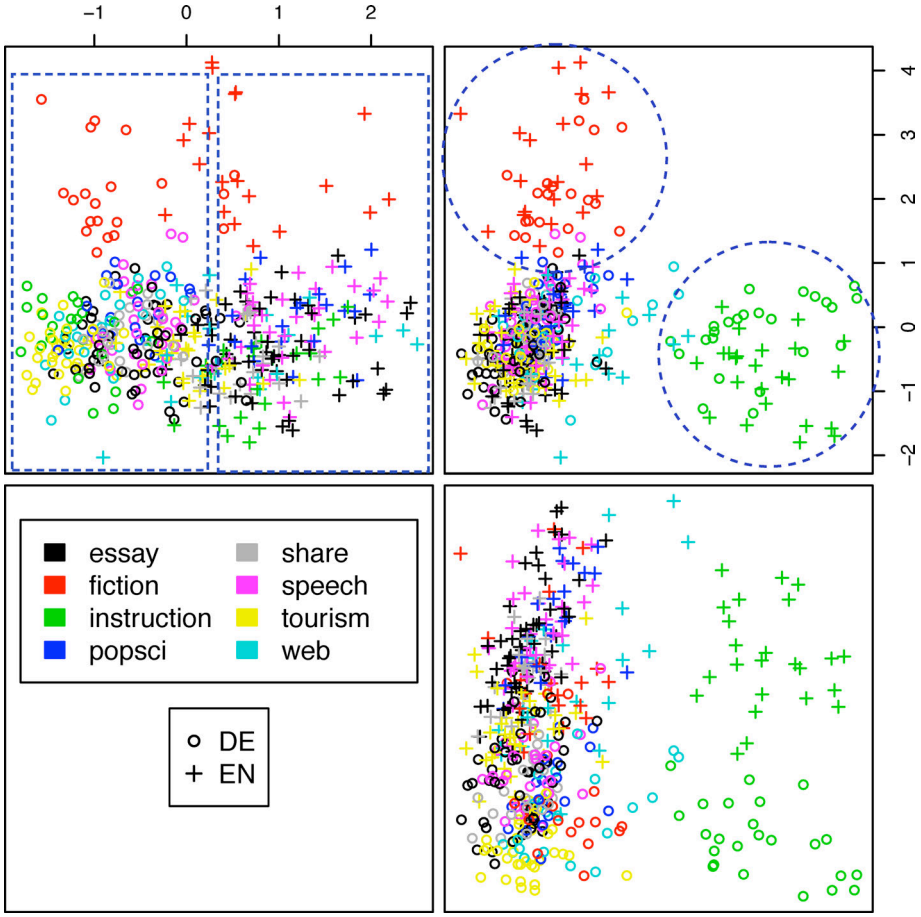

Figure 1. Factor analysis of the data set from case study 1 (regression scores), highlighting language and register of the individual texts.

Apparently, the characteristics of the two registers FICTION and INSTRUCTION MANUAL are similar for English and German. It has to be noted that the 3 latent factors shown in Figure 1 cumulatively account for only 39.1% of the correlations between variables in the data set. However, even when carrying out a higher-dimensional FA and inspecting various combinations of factor dimensions, the other registers can hardly be separated. We experimented with up to 10 factor dimensions, cumulatively accounting for 61.8% of correlations, the maximum we were able to obtain from the unstable FA algorithm.

Our experiments showed that principal component analysis (PCA) yields quite similar results to factor analysis, with the difference that the distinction between English and German does not coincide with a single dimension. The following analyses are based on principal components due to their mathematical and methodological advantages discussed in section 3. The similarity between the scatterplots for factor analysis and principal component analysis (not shown for reasons of space) provides additional support for our decision.

The results obtained so far are interesting for the analysis of register variation as they agree with the qualitative interpretation. Neumann (2008) showed that FICTION and INSTRUCTION MANUAL clearly stand out against all other registers (in both languages), exhibiting a range of similarities mostly with respect to fact orientation. The distinction between originals and translations is less straightforward in terms of a qualitative interpretation. Furthermore, the purely inductive analysis offered by FA and PCA is not sufficient to gain insight into the more complex relationship between translations and originals, thus necessitating the use of semi-supervised methods (see section 4.2).

Step 3 therefore introduces a minimal amount of supervised control in the form of language-external and – as far as possible – theory-neutral categories. In this study, the information provided here is translation status, i.e. whether a text is a non-translated original or a translation (bearing in

mind that the CroCo Corpus contains only texts whose translation has also been included in the corpus). In step 4, linear discriminant analysis (LDA) exploits this information to identify a linear discriminant for translation status, i.e. a weighted combination of features which most clearly separate originals and translations. The distribution of discriminant scores shown in Figure 2 reveals a considerable overlap between originals and translations, indicating that the two categories cannot be clearly separated.

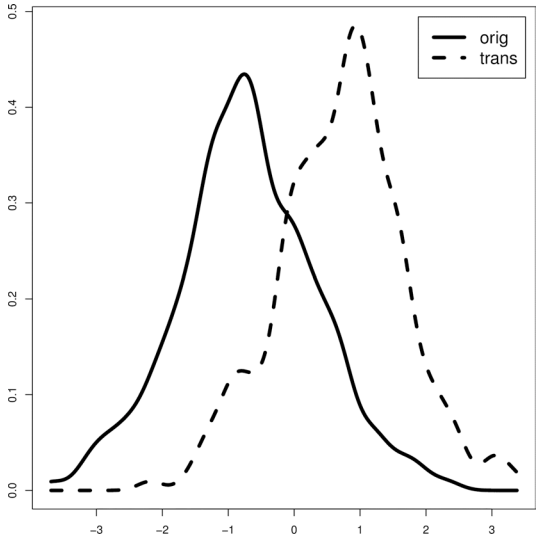

Figure 2. Distribution of discriminant scores for originals and translations (x-axis shows position on the discriminant axis; y-axis indicates how many original / translated texts are near this position, a so-called 'density estimate')

In step 5, the quality of this discriminant is evaluated by leave-one-out cross-validation of the LDA. This approach removes one data point at a time, estimates a discriminant from the remaining data points, and uses this discriminant to predict the category of the excluded data point. Thus, predictions for all data points are obtained while ensuring a clean separation of training and test data in each case. The resulting classification accuracy of 73.1% is significantly better than the baseline of 50% (random choice between translation and original). The selected combination of features thus correlates reasonably well with the differences between translations and originals. However, it fails to characterize the two categories adequately, as shown by the overlap in Figure 2 and the mediocre classification accuracy. According to step 6, we iterate the exploratory procedure with some modifications in order to find a more satisfactory discriminant.

A first indication is the fact that a nonlinear machine learning technique – in particular, a support vector machine with quadratic or radial basis function (RBF) kernel – achieves a much better classification accuracy of approximately 80%. This observation may be explained by the rather intuitive assumption that the characteristic features of translated texts and originals differ between the two languages. We therefore look for separate discriminants in the two languages, by applying linear discriminant analyses separately to German and English texts. The resulting distribution of discriminant scores for each language is shown in Figure 3.

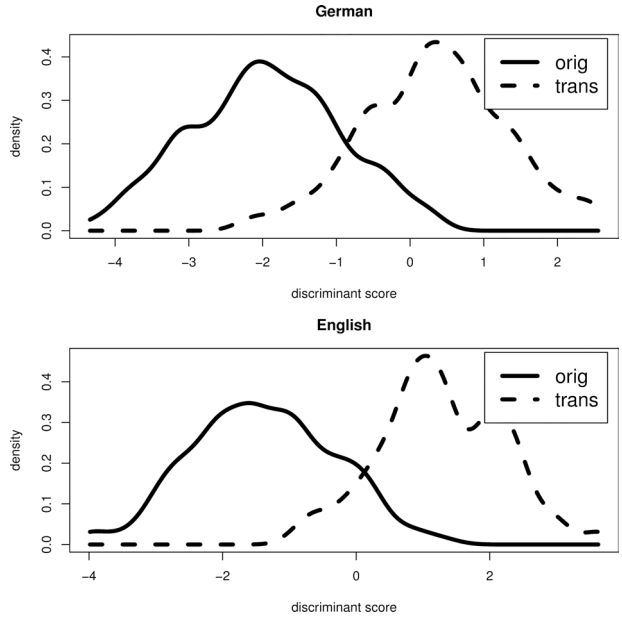

Figure 3. Distribution of discriminant scores for originals and translations in English (bottom panel) and German (top panel).

Cross-validated classification accuracy increases to 80.2% for German and 83.7% for English. Again, non-linear classifiers are even better with an accuracy of more than 85% (SVM with quadratic kernel), but are very difficult to visualize and interpret (cf. Diederich 2008). Having achieved a satisfactory discrimination of the two categories, we can now compute a final set of latent dimensions and proceed with their linguistic interpretation (step 7). For the visualization, it is desirable to combine the discriminant for language and the two discriminants for translation status in a single three-dimensional plot. This is most easily achieved by carrying out a linear discriminant analysis for the cross-classification of language and translation status, in which four categories have to be distinguished: English originals (EO), English translations (ETrans), German originals (GO) and German translations (GTrans). This linear discriminant analysis yields three discriminants shown as a scatterplot matrix in Figure 4. Overall accuracy of the combined LDA predictions for language and translation status is 83.8% according to leave-one-out cross-validation.

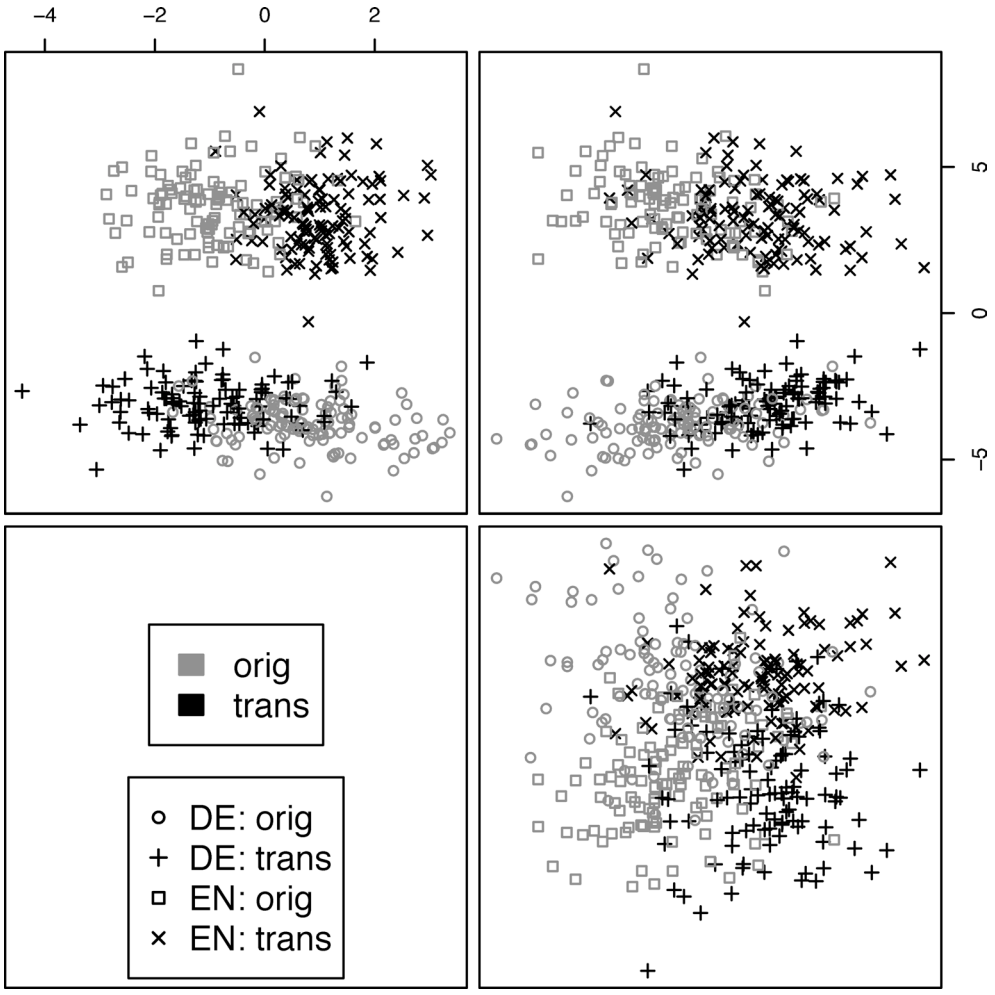

Figure 4. LDA for cross-classification of language and translations status (categories GO, GTrans, EO, ETrans).

The first latent dimension (*y*-axis in the top row) clearly (and trivially) distinguishes the two languages. The second and third latent dimensions allow language-specific discrimination between originals and translations (x-axes in the top row). Remarkably, both dimensions work for German as well as English texts, but in dimension 2 (*x*-axis in top left panel) the German and English translations are on opposite ends of the scale. The linguistic interpretation of these results will be discussed in the following section.

4.2. Discussion

From a linguistic point of view, the multivariate analysis provides a rich and meaningful picture of the variation between originals and translations in English and German. Translation status is – as this study amply shows – a category that requires the application of sophisticated statistical techniques, since the directly observable differences between originals and translations are difficult to interpret. Whereas one feature may point to a tendency to normalize non-canonical features of source texts in the target language, another feature may point to the opposite direction (see Neumann 2008; Hansen-Schirra, Neumann and Steiner to appear). Only the multivariate analysis can shed light on the overall trend in a given translation direction. The following constellations are conceivable.

1. The data set reflects a mixture of originals and translations within the same language, i.e. the impossibility to separate texts according to translation status suggests a clear tendency in translations to normalize features in order to adapt to target language conventions or norms.

2. A separation of originals and translations in the same language points to some kind of translationese (for related bag of words approaches see Baroni and Bernardini 2006; Koppel and Ordan 2011), i.e. language use specific to translations which makes these texts a specific (meta)register.
3. In the case of separation (at least in tendency), translations may either display a tendency to diverge altogether from both source and target texts or they may show the same feature weightings as source language originals. In this case, the specific language use of translations may be attributed to an influence of the source language ("shining through", Teich 2003).
4. If the latter tendency only applies to one translation direction, this may point to an interpretation in terms of diverging prestige of the two languages involved (Toury 1995). More specifically, if the English language is attributed a higher prestige, then this may point to the general role of English as a lingua franca which may have an impact in terms of language contact, corroborating the assumed influence of English on German by translation (Becher, House and Kranich 2009).

In what follows, we will now discuss the three linear discriminant analysis dimensions visualized in Figure 4 and the extent to which they contribute to (i) the attribution of texts to one of the two languages, (ii) the identification of aligned text pairs in both translation directions, i.e. German translations as opposed to their English originals and vice versa and (iii) the identification of translations as opposed to originals in the same language (i.e. the equivalent of comparable corpora).

The linear discriminant analysis first provides a dimension (*y*-axis of the top row) that allows a clear attribution of texts to the two languages. The clearest positive predictors of English in our data set are the verb-related features (finites per sentence, finites per token, infinitives per sentence[8]), nouns and colloquialisms. The strongest negative predictors are pronouns, objects as theme, subordination and adverbials as theme. This means that texts scoring high on the negative predictors can be identified as non-English and consequently as German. This does not necessarily entail that these texts carry all (or any of) the typical predictors of German texts, however: the clear absence of predictors of 'Englishness' is sufficient to identify them as German (since this is the only alternative in this dimension).

The second dimension (*x*-axis of the left column) distinguishes reasonably well between translation pairs: On the left side, the English originals and their German translations can be found, while the German originals and their English translations are to be found on the right side. Interestingly, not all of the registers contribute equally to this classification. In particular the tourism texts deviate from the classification in both languages: All texts from this register tend to the right end of the scale (not shown in the figure). This suggests that the classification accuracy might be even better if register information would be included, too.

It might also be interesting to examine how directly this discriminant reflects the connection between source and target texts. If the individual aligned text pairs – which could be visualized by lines connecting an original and its translation – play an important role for the classification, i.e. if the connecting lines are mostly vertical and parallel to each other, this would point to a strong influence of particularities of the texts and could be attributed to the relatively small size of the corpus. Of course, such an observation would reduce the explanatory power of this classification. If, however, there is no direct relation between the discriminant scores of source and target texts (i.e. if the lines in the visualization do not run in parallel), this could be interpreted as an indicator for a more fundamental difference between non-translated and translated texts in translation pairs (regardless of the concrete language pair).

The contribution of individual features to the classification shows which features predict translation direction. In the translation direction English to German, for instance, the relative absence of subordination (weight –.6193) is the strongest predictor of a text being either an English original or a German translation. By contrast, the relative frequency of objects as themes (weight .5777) is the strongest predictor of German originals or English translations. This is an intuitively plausible predictor, as the relative frequency of the feature in the data set clearly shows (see Table 5). The more flexible German word order permits a reasonable amount of objects in sentence-initial position. This feature is retained to some extent in the translations. By contrast, German translations reflect the low frequency of this

highly marked word order selection in the English originals by displaying a clearly reduced frequency as compared to German originals.

Table 5. Mean relative frequency of object themes in the four subcorpora.

	Object themes per all themes
English originals	0.768%
English translations	1.836%
German originals	9.276%
German translations	3.719%

Furthermore, modal adverbs make a similar contribution to the classification of the translation pair German to English (.5691). Again, the explanation may be found in the area of language contrast. Although German also uses modal verbs to express modality, it is frequently expressed with the help of modal adverbs. The translations into English can be assumed to retain some of the adverbs. Both positive predictors of the translation direction German to English clearly reflect an assumption by Teich (2003) who claims that interference from the source language can be predicted wherever the target language allows a certain option in addition to the more regular option. Our findings show that this may even occur to some extent in features such as object themes that would hardly be assumed to be grammatical at all by introspection. Of course, these marked features may especially facilitate prediction in the multivariate approach.

The third dimension (x-axis of right column) distinguishes between originals and translations, with originals in both languages on the left side and translations in both languages on the right side. This dimension shows somewhat more overlap between the two poles than the other two dimensions, thus corroborating the assumption that translations do not represent a categorically different type of texts outside of the system of the target language. Translations in both languages are best predicted by the frequency of nominalizations (.5324). By contrast, or almost by the same token, the relative absence of nouns (–.8549) predicts originals.

A final aspect that should be mentioned with respect to the linear discriminant analysis plot in Figure 4 is a slight tendency of the German translations to move towards the English pole of the language discriminant, an observation that cannot be made for the English translations. This indicates a more generalized trend towards shining through in the translation direction English to German which may point to the role of prestige mentioned above. The finding is of particular importance because it corroborates and even goes beyond related – and sometimes contradictory – findings based on a conventional analysis of individual features (e.g. Becher, House and Kranich 2009). The tendency in German translations to slightly move towards the English pole of the scale suggests that there may be a long-term trend for German texts to increasingly adopt English features and adjust to the English occurrence frequencies. However, this possibility would have to be further explored in a diachronic study.

The discussion above shows that this case study in principle still allows a direct interpretation of latent dimensions similarly to Biber's multidimensional analysis (e.g. Biber 1995), even though we had to turn to a weakly supervised approach in order to obtain meaningful results. Case study 2 in section 5 will show that a different type of linguistic information purely based on co-occurrences of features does not permit such direct interpretation. The present case study, however, highlights another type of problem resulting from cross-linguistic comparisons. These issues will be discussed in the following section.

4.3. Issues in data-driven cross-linguistic comparisons

Each contrastive study first has to settle on a basis of comparison for the analysis, otherwise its explanatory power is at the very least questionable: formally equivalent features may serve different functions just as seemingly unrelated features may have the same function. This also applies to quanti-

tative cross-linguistic studies: a comparable feature may have different frequencies in the two languages (see Neumann 2011a: 395–396). An inductive study that indiscriminately processes features with a selected statistical technique may therefore produce misleading artifacts. To avoid this, comparable features need to be selected, carefully bearing in mind that the choice does not involve theoretical assumptions about the structure to be discovered. In this case study, this would apply if the features implied the distinction between originals and translations.

We can identify four constellations that need to be addressed by a quantitative study. All four have in common that function is assumed to be the basis of comparison.

1. A feature may be specific to one language and, as a result, produce nonzero frequency counts in one, but not in the other language. A case in point is the German genitive object which does not have a counterpart in English (*Die Gemeinde gedachte der Toten – The congregation commemorated the deceased)*. This constellation exemplifies the trivial indicators addressed in section 2 (cf. Table 4).
2. A contrastive feature may have the same function in both languages and no significant differences in frequency but may still differ in realization. English indirect objects and German dative objects may serve as an example. By and large, both serve the same semantic function, but German dative objects are subject to fewer word order restrictions and therefore do not take a preposition if they are moved to another position as English indirect objects do (*Peter gave the book to Mary – Peter gab das Buch Mary*).
3. A contrastively comparable feature may be realized by the same forms with some clearly diverging usage patterns. In the language pair English-German, modality provides an illustration. In addition to modal verbs (*Karl darf das machen.)*, both languages draw among other features on adverbs to express modality (*They will certainly come back*.). As already mentioned in the previous section, the main contrastive difference here lies in diverging frequency patterns: whereas English can be said to rely more on modal verbs to express modality, German makes more extensive use of modal adverbs.
4. A functionally comparable feature may be realized in part by similar forms, but also by language-specific forms which do not have a direct counterpart in the other language. This can be illustrated by features realizing passive meaning in English and German. In both languages, core passives are realized by a form of *be* or *werden* plus past participle. Alternative realizations in German may have an English equivalent such as the indefinite pronoun *man/one* (even if the frequency patterns again clearly diverge). Modal alternatives such as *sein* + *zu* + infinitive (*Die Behälter sind nach jedem Gebrauch zu verschließen*.) are only available in German. This poses a problem for the multivariate approach: treating language-specific forms as distinct features may again lead to trivial discriminants. Furthermore, frequency patterns for core passives are skewed in contrastive comparison since part of the passive meaning in German is expressed by alternative forms.

Each of these constellations requires a different response by the quantitative linguist. Since the latter case (4.) is the most difficult one, it will be the focus of the following discussion.

Contrastive corpus studies show that combining functional equivalents of voice lead to comparable frequency counts in the language pair English and German. The present study therefore uses the cumulative frequencies of (i) passives with *be/werden* in both languages, (ii) passives with *get* in English[9], (iii) the indefinite pronoun *one/man* in both languages, (iv) the German alternative construction *sein* + *zu* + infinitive and (v) the German alternative *lassen* + *sich* + infinitive. Other marginal forms such as the reflexive verb in German are not taken into consideration here.

An alternative option involves the use of multivariate methods. In order to validate the comparability of features in the two languages, latent dimensions are determined based on the texts from one language only; then all data are visualized according to these dimensions. This was tested for both languages using principal component analysis to determine the latent dimensions. The two registers FICTION and INSTRUCTION MANUAL are reliably separated in both plots. The characteristic features of these registers must be similar in both languages because the German dimensions also separate the English registers, and vice versa. One difference is that the English texts seem to contain more outliers, potentially translations. The German principal component analysis dimensions give a good ac-

count of register variation both in English and German. Neither version of the PCA distinguishes clearly between English and German texts whereas PCA and FA based on all data points does. This suggests that the differences between languages are not situated on the same level as variation within each language.

The discussion of this case study has shown that a weakly supervised multivariate analysis can provide substantial insight into the latent dimensions of linguistic variation. While the 454 texts × 29 features data matrix may still have been amenable to a conventional analysis based on the comparison of individual features, the following section will discuss the application of our methodology to a much larger feature space in case study 2.

5. Case study 2: Non-standard varieties of French

5.1. Multivariate analysis of the data

As described in section 2, case study 2 draws on a data set of French newspaper texts from five African countries as well as France, including two complete volumes of two newspapers for each country. After removing some outliers, the corpus consists of 1,268 individual texts – roughly corresponding to weekly editions of the newspapers – which are analyzed in terms of colligational pairs consisting of a noun and its grammatical function. For practical reasons, only the 18,537 colligational pairs that occur at least $f \geq 500$ times were used as features for the multivariate analysis. The occurrence frequencies of these features in each text were weighted by log-likelihood association scores and scaled logarithmically in order to achieve a more balanced distribution.

In step 1 of our methodological procedure, principal component analysis was employed to identify the most salient 100 latent dimensions. As already pointed out in section 3, factor analysis could not be used because the number of features by far exceeds the number of data points. Note that this particular implementation of our methodology can also be seen as a combination of combinatorial profiles (Blumenthal 2006) with distributional semantics. Figure 5 shows a scatterplot matrix of the first four principal components.

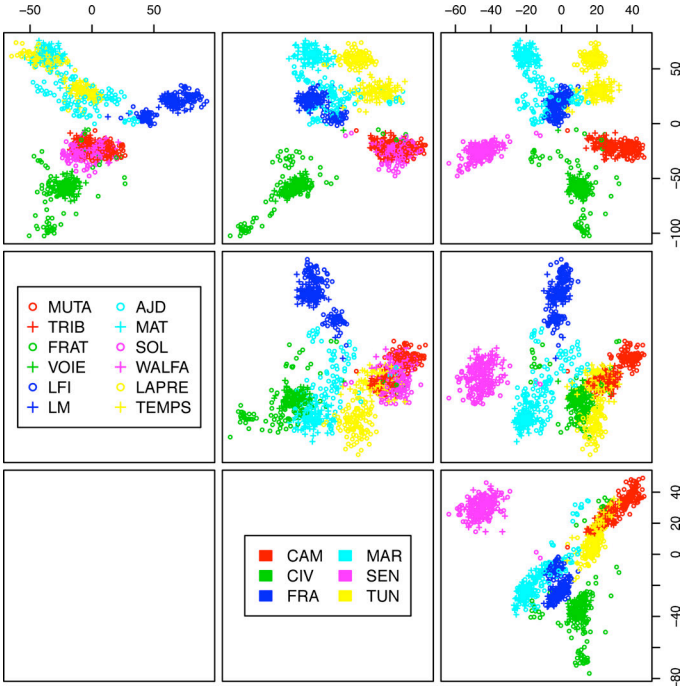

Figure 5. PCA plot of the full data matrix based on 18,537 frequent colligational pairs as features. The scatterplot matrix visualizes the first four principal components.

These four principal components already separate the six countries well (shown in different colors); in many cases, individual newspapers can also be recognized as separate clusters. The visual impression is confirmed by supervised support vector machine classification: both countries and individual newspapers are identified with almost 100% accuracy. At face value, these results seem to suggest that there are clearly distinct national varieties of French at a colligational level. However, a closer look at the principal components reveals a trivial explanation, which holds little interest for linguistic research.

In fact, it does not come as a surprise that the characteristic features of each country include the name of the country and other country-specific proper nouns (for example *Cameroun*, *Yaoundé*, *Biya*[10] for texts from Cameroon and *Sénégal*, *Dakar*, *Wade*[11] for texts from Senegal), as well as artifacts of the newspaper genre such as the abbreviation *AFP*. These are trivial and sufficiently known particularisms (see section 2) which are not at the center of our research interests. Many comparative studies such as keyword analyses are based on such trivial indicators.

Thus, inspection of the visualization (step 2 of our methodology) leads us to modify and optimize the parameters of the analysis. In order to exclude trivial particularisms, we accept as features only shared colligations that occur at least 100 times in the texts from each of the six countries. As a result, the feature space is reduced to 8,248 common dimensions. Note that this feature selection can be seen as a weakly supervised intervention (step 3) since we draw on information about the country of origin of the texts, albeit in a very indirect way.

The principal component analysis dimensions based on shared colligations display an entirely different picture (not shown). The first principal components distinguish neither countries nor individual newspapers, with the exception of a few small groups of outliers.

If there are indeed national varieties, the differences between them are entirely hidden by the language-internal variation across text samples. This impression was confirmed by our exploration of additional principal components: unsupervised hierarchical clustering on 100 principal component analysis dimensions (not shown for reasons of space) completely failed to group together texts from the same country or newspaper.

Following step 3 of our methodology, we now include additional knowledge for a weakly supervised search of relevant discriminants. In this case study, the identities of individual newspapers are used as theory-neutral categories. Crucially, we do not indicate the respective country of origin in order to avoid making assumptions about the existence of national varieties of French. A linear discrimi-

nant analysis for the 12 newspapers as categories yields 11 discriminants, with almost perfect classification accuracy of 97.6% (leave-one-out cross-validation). Surprisingly, the first six discriminants are already sufficient to distinguish all newspapers; and even the first four discriminants visualized in Figure 6 separate all but the two Moroccan newspapers clearly.

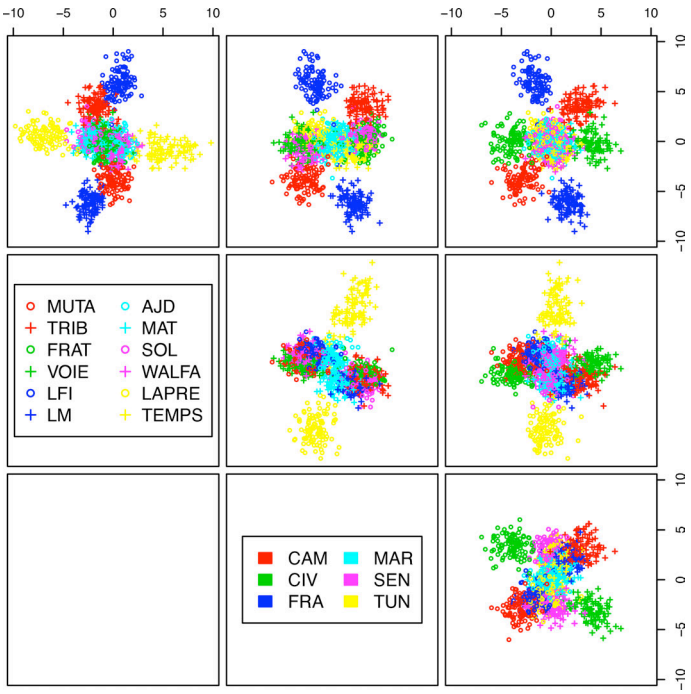

Figure 6. Scatterplot matrix of the first four LDA dimensions (shared colligations).

The newspapers from Cameroon, Tunisia, France and Cote d'Ivoire are most clearly distinguished by the first four discriminants. It is particularly remarkable that the two newspapers of each country can be found in opposing positions along a single axis, while all other newspapers take a more or less neutral position on this axis. Note that the linear discriminant analysis algorithm did not 'know' that these pairs of newspapers are from the same country and therefore could, for instance, equally well have placed two newspapers from different countries on the same axis. Or, even more likely, could have singled out each newspaper on a separate discriminant dimension.

The axes, which can now be used to identify each country with high accuracy, do not coincide with the linear discriminant analysis dimensions, but they can easily be identified in the scatterplot matrix. This again emphasizes the importance of visualization (step 2 of our methodology). For instance, the first row of Figure 6 shows the axes for Tunisia and France as well as somewhat less clearly for Cameroon. The bottom right panel shows an axis for Cote d'Ivoire.

This surprising result implies that country-specific varieties indeed emerge inductively from the data, more specifically from the variation between individual newspapers. However, the varieties are not so much characterized by frequent (or particularly rare) colligations, but rather by contrasting options which distinguish the two papers from the same country. While this could in principle be a coincidental pattern in the high-dimensional feature space, the likelihood of newspapers consolidating in this way by chance is extremely low.

In order to determine the shared colligations that contribute to country-specific varieties of French, we 'automate' the visual identification of axes by carrying out a fully supervised linear discriminant analysis in the six-dimensional discriminant space. For each country, linear discriminant analysis is used to distinguish the two newspapers, ignoring all other data points. Note that this does not invalidate our inductive approach, since the supervised linear discriminant analysis is only used to determine axes of variation that are already visible in the scatterplot matrix (Figure 6).

The distribution of discriminant scores on the resulting six axes (Figure 7) confirms our visual impression that the two papers from the same country are arranged on the opposing positions whereas all other newspapers remain largely neutral. Quantitative evaluation with a support vector machine shows an almost perfect country identification accuracy of more than 98.7% on the first six latent discriminants (using a quadratic kernel so the two poles of each axis can be identified), and 91.7% using only the four discriminants visualized in Figure 6 (accuracy determined by ten-fold cross-validation).

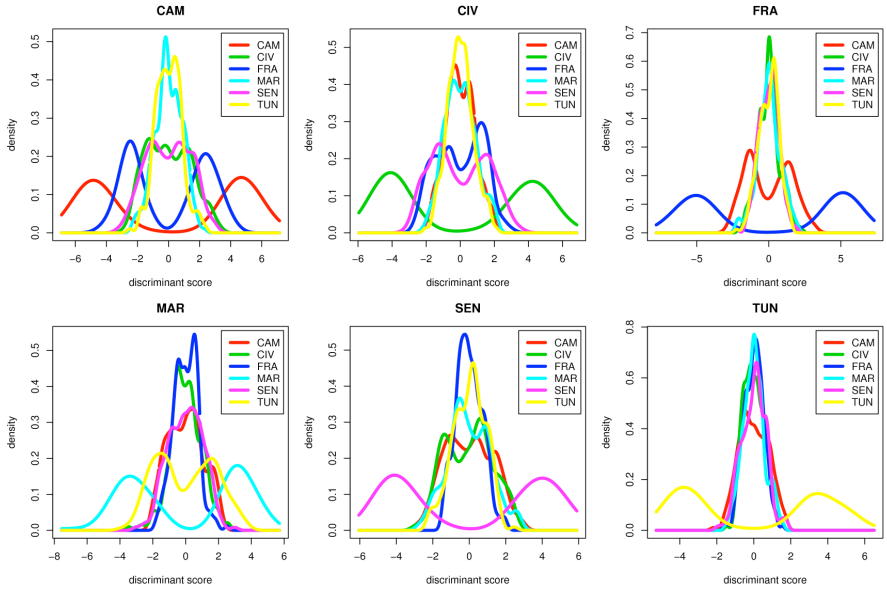

Figure 7. Distribution of discriminant scores on visually identified axes: the two newspapers from each country are located on opposing poles of the respective axis, while most other newspapers occupy neutral locations.

In order to exclude the possibility that these astonishing results are merely an artifact of the high-dimensional data and the analysis procedure, further experiments were carried out. Newspapers were brought into a randomized order without any change in the results. Furthermore, the linear discriminant analysis algorithm was applied to the 24 individual volumes (two volumes from each of the 12 newspapers) as categories. This, too, only led to minor changes. The classification into country-specific axes with two poles thus appears to be stable. A further explanation for these axes can only be derived from a linguistic interpretation of the underlying colligational pairs. This will be attempted in the following section.

5.2. Discussion

The reduced set of shared colligations reveals some clear functional restrictions for specific nouns with respect to several country-related samples. Thus, taking into account the calculated discriminant weights, for example, the association of *accent* and *confiance* with the function of Direct Object and the association of *évidence* with the function of Prepositional Complement appear to be well discriminating features within the samples representing Cameroon (*accent*, *évidence*), Cote d'Ivoire (*confiance*), Senegal (*évidence*) and Tunisia (*accent*). The distinctive potential of these colligational associations is due to the fact that in one of each newspaper making up the corresponding sub-samples, the nouns in question are characterized by a highly stereotyped use confined to the Verb-Noun phrasemes *mettre l'accent sur qc.* ('to put the emphasis on sth.'), *faire confiance à qn.* ('to trust sb.') and *mettre qc. en évidence* ('to bring sth. to light', 'to highlight sth.').

At the same time, the results obtained for the discriminants of all shared colligations show that even with this criterion – introduced to exclude country-specific thematic particularities (mainly the name of the country itself and other named entities from this country) – such effects cannot be completely prevented (see below for a more detailed discussion). This should not come as a surprise, considering that nouns are the most content-bearing and entity-related lexical category. It also indicates

that experimental work based on noun colligations presents a considerable challenge to our inductive methodology, which has to deal with a mixture of effects in an appropriate way.

The axis distinguishing the two Tunisian newspapers is a good case in point. If we take the 20 colligations that are the best indicators for each newspaper (i.e. have the largest weights in the discriminant), two small sets of noun lexemes appear, some of which are restricted to only one functional association (see below *instance*, *solidarité*), while others occur in several colligations. These lexemes indicate different thematic orientations of the two newspapers, viz.

1. reports on legal and administrational topics with the lexical set *accusé*, *instance*, *agent*, *tribunal* for the newspaper *Le Temps*

and

2. presentation of governmental activities with the items *Ali*[12], *ministre*, *secrétaire*, *solidarité*[13] for the newspaper *La Presse*

On the other hand, we can observe that a single item – the abbreviation *M.* of the polite form of address *Monsieur* (as well as its female equivalent *Mme* for *Madame*) – is the single most distinctive feature for three other countries (France and, to a lesser degree, Cameroon and Cote d'Ivoire), again irrespective of its colligational patterns. The use of this item by itself obviously accounts for clear differences between newspapers with respect to their stylistic conventions, but it does not hint at far-reaching systemic choices that represent different national varieties (although it is not an indication of thematic effects either).

Note, however, that individual lexical items such as the abbreviation *M.* cannot be the only distinctive factor e.g. for newspapers from France. If this were the case, the newspapers from Cameroon and Cote d'Ivoire would also be shifted towards the poles in the top right panel of Figure 7 (because *M.* is also a highly distinctive feature for these countries). Discriminants obtained from a high-dimensional multivariate analysis are typically based on combinations of many dozens or hundreds of individual features. On the one hand, this is an advantage of multivariate methods over more conventional approaches; on the other hand, the linguistic interpretation of such discriminants is a difficult and time-consuming task.

To sum up and with the guiding questions for this case study from section 2 in mind, our methodological approach still needs further refinement in order to be applied to a successful colligational analysis of regional variation. Although it does seem to be possible to discover country-related differences by means of a weakly supervised statistical exploration of the corpus data, the individual features we have been able to single out do not provide direct insight into divergent systemic choices that can be attributed to the existence of divergent regional varieties. They rather seem to reflect thematic effects (cf. the example of Tunisia above) or language conventions observed by a small, clear-cut community of practice, viz. the editorial staff of a particular newspaper (cf. the case of *M.* in France, Cameroon and Cote d'Ivoire). The data and parameters chosen for this case study – working with single colligational items and low-level abstractions in the annotation – play their part in bringing about the mixed effects we have observed.

6. Conclusion and outlook

This paper introduced an approach to the corpus-driven, quantitative analysis of linguistic variation that extends widely-used multivariate methods with extensive visualization and weakly supervised exploration of high-dimensional feature spaces.

We demonstrated that this general methodology – albeit with some adaptations in parameter settings and statistical techniques – can be used for the investigation of two quite diverging data sets in different languages. Moreover, it was shown that visualization is essential both for the targeted exploration of latent dimensions and for the linguistic interpretation of the results. The experiments reported

in sections 4 and 5 revealed that much of the latent information hidden in the data can only be discovered with the help of a semi-supervised exploration of the results.

The linguistic discussion of the case studies showed that the moderate number of lexicogrammatical features (i.e. relatively high-level abstractions) in case study 1 permit a more straightforward interpretation of the results of the multivariate analysis. It was still necessary to avoid non-comparable features that would lead to a trivial distinction between English and German texts. The semi-supervised approach provided some very interesting findings: As far as register variation is concerned, the multivariate analysis corroborated the qualitative interpretation of feature combinations in Neumann (2008). In terms of translation status, the latent discriminant analysis suggests some highly interesting quantitative findings which will certainly help to advance empirical translation research in particular in terms of the impact of one language on another shown here for English and German. Future work in this area will involve more fine-grained analyses of different data combinations. It appears particularly promising to concentrate on individual registers and their contribution to the overall distinction between translations and originals.

The colligational approach in case study 2 – based on a large number of relatively low-level features – led to a considerable number of direct indicators of variational classification (such as named entities and stylistic conventions), which may be of great interest to cultural or media studies. The linguistic interpretation of these indicators proves more problematic. Future work in this area therefore involves the selection of colligations that are not connected to named entities and other direct indicators, in order to focus on the characteristic linguistic properties of national varieties of French. We will also explore statistical techniques for the exclusion of direct indicators, for the identification of groups of less prominent, but related features that contribute to the multivariate discriminants, and for the separation of lexical and colligational effects.

On the methodological side, future work concentrates on further case studies to validate the procedure outlined in section 3, as well as on the exploration of additional statistical techniques (e.g. probabilistic latent class models). If future experiments corroborate the promising results of this paper, our approach may be established as an inductive multivariate methodology that is able to discover fine-grained latent information in linguistic data. The divergent nature of our case studies suggest that it can be used both for dialectological (and other variational) data and for more typologically oriented data.

As to situating this paper vis-à-vis dialectological and typological research, the dialectological perspective is clearly addressed by our case study 2. Case study 1 with its cross-linguistic design might help extend typological research to such fine-grained explorations of the relatively subtle differences between closely related languages. Yet, it does not seem applicable to larger-scale typological comparisons since the features underlying the vector spaces need to be comparable in the sense of representing a functionally similar phenomenon: both languages – this is an essential aspect of our approach – are analyzed in a common vector space.

Acknowledgments

The authors would like to acknowledge partial support by the German Research Foundation on case study 1 in the CroCo project (DFG project no. STE 840/5-1, STE840/5-2 and HA 5457/1-2). The authors wish to thank the two reviewers for their valuable comments which helped to clarify the main points of the paper.

Toolbox

- multivariate latent variable models: factor analysis (FA), principal component analysis (PCA)
- visualization: scatterplot matrix, distribution of factor/dimension scores
- weakly supervised analysis: linear discriminant analysis (LDA), evaluated in terms of classification accuracy (leave-one-out cross-validation)

- supervised machine learning for quantitative evaluation: support vector machines (SVM) with linear and quadratic kernels, evaluated by ten-fold cross-validation

References

Baroni, Marco and Silvia Bernardini 2006 A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing* 21 (3): 259–274.

Becher, Viktor, Juliane House and Svenja Kranich 2009 Convergence and Divergence of communicative norms through language contact in translation. In: Kurt Braunmüller and Juliane House *(eds.), Convergence and Divergence in Language Contact Situations*, 125–152. Amsterdam: Benjamins.

Benzakour, Fouzia, Driss Gaadi and Ambroise Queffélec 2000 *Le français au Maroc: lexique et contacts de langue*. Louvain-la-Neuve: Duculot.

Biber, Douglas 1988 *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

----. 1995 *Dimensions of register variation*. Cambridge/New York: Cambridge University Press.

Blumenthal, Peter 2006 *Wortprofil im Französischen*. Tübingen: Niemeyer.

Diederich, Joachim (ed.) 2008 *Rule Extraction from Support Vector Machines*, volume 80 of *Studies in Computational Intelligence*. Springer, Berlin, Heidelberg.

Halliday, Michael A. K., and Ruqaiya Hasan 1989 *Language, context, and text: aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, Michael A. K. 2001 Literacy and linguistics: Relationships between spoken and written language. In: Anne Burns and Caroline Coffin (eds.), *Analysing English in a global context*, 181–193. London: Routledge.

Hansen-Schirra, Silvia, Stella Neumann and Erich Steiner to appear *Cross-linguistic Corpora for the Study of Translations - Insights from the language pair English-German*. Berlin: de Gruyter Mouton.

Hoey, Michael 1995 *Lexical Priming: a new theory of words and language*. London: Routledge.

Koppel, Moshe and Noam Ordan 2011 Translationese and its dialects. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326. Portland, Oregon: Association for Computational Linguistics.

Martin, J.R. 1992 *English text*. Amsterdam: Benjamins.

Matthiessen, Christian M. I.M. 1993 Register in the round: diversity in a unified theory of register analysis. In: Mohsen Ghadessy (ed.), *Register analysis. Theory and practice*, 221–292. London: Pinter.

Neumann, Stella 2008 *Contrastive register variation. A quantitative approach to the comparison of English and German*. Unpublished Habilitationsschrift. Saarbrücken: Universität des Saarlandes. (to appear in the Trends in Linguistics. Studies and Monographs series, de Gruyter)

Neumann, Stella 2011a Contrasting frequency variation of grammatical features. In: Marek Konopka, Jacqueline Kubczak, Christian Mair, František Štícha and Ulrich H. Waßner (eds.), *Grammatik und Korpora 2009: Dritte Internationale Konferenz*, 389–410. Tübingen: Narr.

Neumann, Stella 2011b Assessing the impact of translations on English-German language contact: Some methodological considerations. In: Svenja Kranich, Viktor Becher, Steffen Höder, and Juliane House (eds.), *Multilingual Discourse Production. Diachronic and Synchronic Perspectives*, 233-256. Amsterdam/Philadelphia: Benjamins.

Niangouna, Augustin and Ambroise Queffélec 1990 *Le français au Congo (R. P. C.)*. Aix-en-Provence: Publications de l'Université de Provence.

Nzesse, Ladislas 2009 *Le français au Cameroun : d'une crise sociopolitique à la vitalité de la langue française (1990-2008)*. Nice: Institut de Linguistique Française [= *Le français en Afrique* 24 (2009), Revue des Réseau des Observatoires du Français Contemporain en Afrique].

Queffélec, Ambroise 1997 *Le français en Centrafrique : lexique et société*. Vanves: Editions Classiques d'Expression Française (EDICEF).

Schneider, Edgar. W. 2007 *Postcolonial English: varieties around the world*. Cambridge: Cambridge University Press.

Stein, Achim 2003 Lexikalische Kookkurrenz im afrikanischen Französisch. *Zeitschrift für französische Sprach- und Literaturwissenschaft* 113: 1–17.

Tapanainen, Pasi and Timo Järvinen 1997 A non-projective dependency parser. *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, 64–74.

Teich, Elke 2003 *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Berlin/New York: Mouton de Gruyter.

Toury, Gideon 1995 *Descriptive translation studies and beyond*. Amsterdam: Benjamins.

Notes

1. For the purposes of the present study eight texts with incomplete annotation were excluded.
2. Cf. Niangouna & Queffélec (1990), Queffélec (1997), Benzakour & Gaadi & Queffélec (2000), Nzesse (2009) to mention just a few examples of a whole series of contributions.
3. As shown by Schneider (2007:46f) with respect to post-colonial varieties of English, the existence of deviating syntagmatic patterns provides an essential indication of emerging endogenetic norms.
4. For a discussion of the notion of 'combinatorial profile', see Blumenthal (2006) amongst others.
5. The automatic corpus annotation was carried out with a commercially licensed parser developed by Connexor (cf. Tapanainen & Järvinen 1997).
6. One volume of the newspaper *Notre Voie* has only partially been included in the corpus.
7. Every dependency relation is considered as passive nominal valency which holds between a noun and its verbal, adjectival or nominal head with respect to which it is in the functional syntactic position of either Subject, Object, Complement and Adjunct or Modifier. Every relation is considered as active nominal valency which holds between a noun as head and its determiners, modifiers and complements.
8. The verb-related features are a good example of strongly correlated features, as discussed in section 4.1. Since these features co-vary, the interpretation should not focus too much on the fact that all three features receive high LDA weights.
9. The expected skewedness for different registers cannot be taken into consideration in this approach.
10. Name of the current President of Cameroon
11. Name of the current President of Senegal.
12. Name of the former President of Tunisia.
13. Used as postmodifier to the nouns *ministre*, *fonds* or *banque* in compound named entities such as *ministre des Affaires sociales et de la Solidarité*, *Fonds de Solidarité Nationale* and *Banque Tunisienne de Solidarité*.