RECHERCHES UNIVERSITAIRES
FLSHS

ISSN 2811-6585

Academic Research

مجلة بحوث جامعية

Revue d'études littéraires, linguistiques et de sciences humaines

N° **19** Vol. **1** (*April* 2025)

Special Issue

Editors

Akila Sellami-Baklouti
Fatma Benelhaj
Sabiha Choura
Nadia Abid

This special issue compiles papers from the 48th Systemic Functional Congress (ISFC48 organised in March 2023 by the Systemic Functional Linguistics Association of Tunisia (SYFLAT) and the Laboratory of Approaches to Discourse (LAD-LR13ES15), under the auspices of the Faculty of Letters and Humanities at the University of Sfax. This special issue, which explores the theme of power and empowerment in relation to language and systemic functional theory, is divided into two volumes. The contributions in this first volume provide some reflections on SFL notions, which can empower both the theoretical apparatus and its application to different types of discourse. The papers in the second volume showcase how SFL language descriptions can empower pedagogical practices.

https://recherches-universitaires-flshs.com

Ce site permettra aux internautes qui s'y inscriront via l'<<Espace Membre>> de consulter ou de télécharger des articles déjà parus dans les numéros précédents de la revue ou alors de soumettre des articles pour évaluation à paraître après acceptation dans un prochain numéro.

RECHERCHES
UNIVERSITAIRES
FLSHS

ISSN 2811-6585

Academic Research

مجلة بحوث جامعية

Acknowledgements

The editors would like to express their sincerest thanks to the esteemed scholars who kindly contributed to the review process in this special issue.

- Radhia Bebes
- Maria Brisk
- Izaskun Elorza
- Najla Fki
- Lise Fontaine
- Sondes Hamdi
- Ameni Hlioui
- Imen Ktari
- Fayssal Maalej
- Anne McCabe
- Mimoun Melliti
- Dorra Moalla
- Nesrine Triki
- Zhenhua Wang

The editors would also like to express their sincere gratitude to Prof. Sadok Damak, Editor-in-Chief of *Academic Research,* for his invaluable support and guidance throughout the preparation and publication of this special issue.

Table of Contents

Multivariate exploration of instantial variation in situational context: The powerful role of the individual instance of language use

Stella Neumann

Abstract

Register variation is usually discussed in terms of variation between groups of texts as instances of a given register, i.e. as between-group variation. This suggests that different contexts represent clear-cut categories, thus leaving open the question of how to conceptualise texts representing borderline instances and instances that show traits from two categories. In this paper, I argue that viewing registers as categories obfuscates continuities between textual instances across categories reflected in similar distributions of linguistic features. To demonstrate the effect, I will visually explore results of the quantitative multivariate analysis of data from the International Corpus of English. The representation of every instance in visualisations of the results yielded by multivariate analysis reveals widespread continuities and a substantial amount of overlap between register-based groups of texts. This 'fuzzy' character of registers is understated by empirical studies of register variation, which average out textual instances on the basis of register labels. The paper shows that systemic functional linguistics has a powerful conceptual toolkit to capture the fuzziness of register variation and that it can be explored with the help of Geometric Multivariate Analysis.

Keywords

Instantiation; topology; geometric multivariate analysis; situational context; quantitative corpus study

1. The Concept of Instantiation

Each time a language user expresses meaning with the help of language, they make myriads of choices from the language system. Each one of these choices has an impact on the shape of the system by perturbing the probability of the chosen feature within the system. While a single instance of perturbing the system will go unnoticed in the uncountable numbers of instances produced on any single day in a language, making the same choice repeatedly can result in a trend within the language system. This choice can take the form of reinforcing one existing option instead of alternative options or giving prominence to a novel option.

Choosing can also be seen as an activity of a whole population, when the individual citizens vote in an election (Halliday 2013, 22). In this context, too, the individual vote itself remains invisible, sometimes giving citizens the impression that it does not make a difference whether they vote or not. But, just like the accumulation of individual votes can give power to a particular candidate, the accumulation of individual choices by many language users can change the overall language system. So, the individual language user also has power in shaping their language. This paper addresses the role of the individual instance - from the perspective of the individual text - within the overall system. Systemic functional linguistics offers a theoretical concept that captures the inextricable link between system and instance: the concept of instantiation.

Halliday & Matthiessen (2014, 27) characterise the language system as a meaning-making potential and the individual text as an instance of this underlying system. The textual instance can be viewed as a manifestation of the language in which specific choices have been made from the options available in the system. This relationship between system and text is viewed as a cline, "a continuum carrying potentially infinite gradation" (Halliday 1961, 249). As Halliday (1999, 9) argues using his well-known metaphorical comparison to the relationship between climate and weather, we cannot identify a cut-off point between an overall (stable) regularity captured in the system and a transient pattern. Likewise there is no way of determining whether variation between two texts is simply individual variation or whether this variation actually reflects a more systematic pattern.[1] In between the two poles of the cline, we can therefore identify groups of instances that share certain similarities as text types, or, when starting from the system end, these types can be regarded as registers (Halliday & Matthiessen 2014, 29).

Instantiation clearly foregrounds the relevance of textual instances. Nevertheless, there is some tension between how constructs are conceptualised and how we refer to them – and

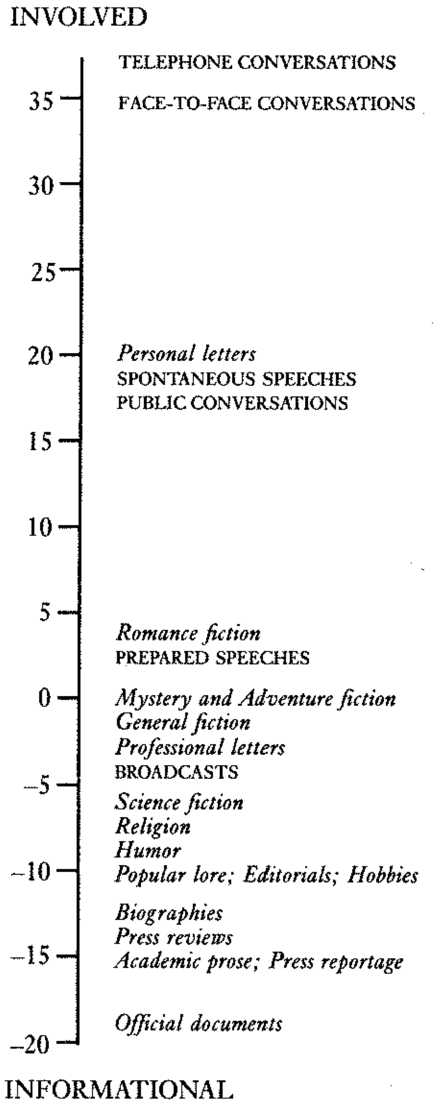
[1] Given the methodological inroads made in quantitative linguistics during the past 25 years and the probabilistic view of language built into Halliday's theorising gaining ground in various branches and schools of linguistics (see, e.g., Beckner et al. 2009; Bod et al. 2003; Grafmiller et al. 2018), we might actually be coming close to determining such cut-off points.

specifically how we investigate them methodologically. When we investigate groups of texts, we tend to refer to them by culturally established labels (Biber et al. 2021, 20) such as recipes, weather forecasts, rental agreements, e-mail messages, inaugural speeches (see Halliday & Matthiessen 2014, 29 who refer to these as text types). Such labels imply that individual texts can be categorised as instances of a register, as in or out with respect to a category, effectively treating the categories as discrete. More specifically, register variation is usually discussed in terms of variation *between* groups of texts. This suggests that we can identify boundaries between registers, even if we accept the idea that within a register there is variation between instances.

Systemic functional research is particularly strong on conceptualising instantiation and the role of variation between instances. However, work validating this conceptualisation with the necessary large data sets is scarce. Studies of small corpora cannot provide a reliable assessment of the range of variation (see also Biber et al. 2021, 25). Biber's quantitative multidimensional analysis approach (MDA; e.g., Biber 1995), which is not affiliated with any particular theoretical framework, can reasonably be called the mainstream approach to register variation and has brought about important insight into actual linguistic patterns by applying the multivariate technique of factor analysis to large corpora. Factor analysis is used to discover latent dimensions of shared variance in the data set. However, one shortcoming of quantitative register studies (including MDA) is their reliance on interpreting aggregated scores for entire registers along the dimensions of variation.

This issue can be illustrated with the first dimension of Biber's (1995, 146) MDA reproduced here as Figure 1. Each register is represented by an identifiable position along a dimension based on the mean score of all texts labelled as belonging to this register. Although Biber (1989, 41) addresses overlaps between the registers, the visualisation does not reflect any overlaps but suggests rather neatly separated categories. Naturally, variation between instances is captured by the mean, but in the interpretation of the data it is lost.

Figure 1. Mean scores of dimension 1 for 23 English registers yielded by factor analysis



(Source: Biber 1995, 146)

This loss of information of fundamental theoretical interest is the result of data aggregation by latent variables, that is, by meta information about the texts. Grouping data by variables such as register labels runs the risk of overstating the differences between categories and understating variation within a category. This has been demonstrated by Brezina and Meyerhoff (2014) for social variables. They compared linguistic indicators claimed by previous corpus-based sociolinguistic studies to be associated with social categories in statistically significant ways with the same indicators in a suite of random groupings not capturing any specific social variable. This randomly aggregated data yielded statistically significant results for several socially meaningless groupings, leading Brezina and Meyerhoff to conclude that the findings of previous studies are more likely due to the (aggregative) method rather than substantive sociolinguistic variation. Moreover, they also show that accounting for variation between individual speakers in the statistical analysis yields more reliable results. Brezina and Meyerhoff therefore advocate considering individual variation in addition to social variation. One of the steps they recommend for accounting for individual variation in the data is to visually inspect data with scatter plots (Brezina & Meyerhoff 2014, 24). This type of diagram represents individual data points in the form of symbols in two-dimensional space based on their values for two (numerical) variables. Using different colours and shapes, grouping information can be included. In this way, scatter plots allow the researcher to explore meta information, e.g. social variables, to account for patterns in the data, while not losing sight of continuities and overlaps between groups by representing each data point along the two dimensions of the plot.

This paper sets out to examine the contribution of individual instances to emerging registerial patterns with the help of such scatter plots representing results of the multivariate analysis of register variation across varieties of English reported by Neumann and Evert (2021). These scatter plots form the interpretive backbone of Geometric Multivariate Analysis (GMA) developed by Diwersy et al. (2014), Evert and Neumann (2017) and Neumann and Evert (2021). As will be explained in more detail in section 2, they visually represent the scores of each individual data point for the dimensions yielded by the multivariate analysis.

After a conceptual overview of GMA in the next section, I will introduce the data used for illustration in this paper (Section 3). The following section “Making Sense of Instantial Variation” is dedicated to discussing the topological perspective on variation between textual instances brought out by the multivariate analysis. The paper closes with some concluding remarks.

2. Geometric Multivariate Analysis

There is ample empirical evidence showing how ensembles of linguistic features co-vary in response to external factors such as situational context, social variation etc. (e.g. Biber 1995; Diwersy et al. 2014; Kruger & Van Rooy 2016; Xiao 2009). GMA can be thought of as an alternative to Biber's multidimensional approach to the analysis of linguistic variation. Whereas MDA focuses on revealing correlation patterns between co-occurring linguistic features with the help of factor analysis, GMA is primarily concentrated on discovering differences between texts that allow the analyst to identify distinct clusters of texts (Neumann & Evert 2021, 148). These clusters are explored with a suite of multivariate analysis techniques involving mainly principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is an unsupervised technique, i.e. one that computes patterns of co-occurring features without having any information on the categories that might explain these patterns. Based on these patterns, it reveals the hidden structure of a data set, which is assumed to be driven by the above-mentioned external factors.

Since PCA only exploits differences between data points in terms of feature frequencies, it gives a general overview of tendencies and cannot capture subtler patterns in the data. GMA therefore complements PCA with a related machine learning technique, namely LDA, which brings out more fine-grained patterns in the data. It does so by exploiting information about a category which applies to the data points and is introduced to the algorithm. It is thus a supervised technique. Neumann & Evert (2021), for example, included information on the specific text category with which each text was labelled (see below). This information allows the LDA to find a perspective on the data that minimises differences between data points *within* groups (i.e. all texts belonging to one category), while, at the same time, maximising the differences *between* groups. The choice of these techniques for this particular statistical analysis is justified by their mathematical properties (Diwersy et al. 2014). Crucially, both represent data points, i.e. texts, spatially along dimensions which capture the complex constellation of similarities and differences between data points. The geometric distances between data points represented by PCA and LDA are then interpreted as linguistic (dis)similarities between texts. This, in turn, facilitates the visual interpretation of the configuration of the data set (see below). GMA foregrounds visualisation as a means to explore the precise positioning of individual texts in relation to all other texts. In this way, it accomplishes what Brezina and Meyerhoff (2014) suggest, namely to account for group-related and individual variation simultaneously.

Given these characteristics, GMA is conceptually designed to align with systemic functional language theory, specifically with aspects of instantiation. The selection of linguistic features for the analysis based on theoretical considerations is also warranted by the data-driven character of the analysis.

The goal of the study of linguistic variation is to understand differences between instances of language use, i.e. texts, and to what extent they form clusters by certain factors. Variation between texts is likely to affect multiple features across all strata. The effect is thus complex, reflecting multivariate relationships between features (and factors). Studies of linguistic variation are therefore well advised to include an entire set of linguistic features and to account for the complex interaction between them. To ensure capturing the widest possible range of variation, GMA uses a large feature set based on register theory. In systemic functional linguistics, register represents the central organising principle of variation along the cline of instantiation. The contextual variables grouped under the parameters of field, tenor and mode (Halliday & Hasan 1989) can be seen as the highest level filter within a given (variety of a) language (Matthiessen 1993). Out of the possible options available in a language system, they constrain the options available for choice in a given situation, or, more often, they condition the probabilities of options to be chosen. Using a set of linguistic features that captures the (probabilities of) linguistic choices across registers is therefore suitable to characterise texts comprehensively for various types of linguistic variation.

To this end, GMA draws on “queryable” versions of features originally operationalised by Neumann (2014) for quantitative analysis of register from systemic function register theory and then represented as corpus queries (see below). Field is captured by indicators such as the distribution of nouns, nominalisations, neoclassical compounds, attributive adjectives and lexical density. Tenor is reflected in indicators such as first, second and third person pronouns, interrogatives and imperatives, modal verbs, titles and forms of address. Mode is represented by indicators such as coordinating and subordinating clauses, place and time adverbs and an approximation of theme.

Counting many features in a large corpus requires computational support. Neumann and Evert (2021) use a computational script that automatically extracts frequency information from a corpus annotated with part-of-speech information with the help of the IMS Open Corpus Workbench, a powerful computational tool set for managing and querying large corpora (Evert & Hardie 2011). More specifically, the script executes pre-defined queries for all features, counts all occurrences and writes all counts by text in a table. These feature counts are then normalised for meaningful units of measurement such as the number of nouns as a proportion of all words in the text, the number of passives as a proportion of all finite verbs (as a proxy for the grammatical clause) in the text, or the number of words as a proportion of all sentences in the text. Moreover, to balance out the magnitude of difference between the different feature frequencies, they are standardised as z-scores and, where useful, additionally log-transformed to mitigate the influence of far outliers. For each data point, meta information such as the text category assigned to the text, the variety or language of the text, etc. is collected. The meta information is later used to explore potential clusters of data points in the visualisation of the spatial configuration of data points. One of these metadata categories is used as input for the LDA.

Each text in the corpus is thus represented as a data point in the multivariate analysis, characterised by the (standardised and transformed) frequencies of all features. In technical parlance, such a data point may be called a feature vector, which can be represented in feature space.

3. Neumann and Evert's Data Set

I will exemplify the exploration of instantial variation in GMA with the help of a study of register variation across three varieties of English by Neumann & Evert (2021).[2] They expected to observe differences in comparable registers across varieties of English because the different regions of the world in which English is widely spoken are likely to differ culturally and situational context cannot be divorced from the cultural context in which situated interaction occurs. They chose Hong Kong English (HKE), Jamaican English (JAE) and New Zealand English (NZE), three varieties spoken on different continents and characterised by different historical developments. Moreover, these varieties represent different language contact situations, namely in the case of NZE primarily L1 use, in the case of JAE an indigenised L2 variety (its label in the Electronic World Atlas of Varieties of English; Kortmann et al. 2020), and in the case of HKE an L2 variety that competes with predominant use of another L1 by its users. Usefully, components of the International Corpus of English (Greenbaum 1996) exist for these varieties. This corpus offers 1 million word corpora of varieties of English based on a common corpus design including 20 different text categories intended to capture various spoken and written registers of English. It should be noted that the language use targeted by the compilers is educated English. As a consequence, the corpus does not capture the full range of conceivable variation within each variety. Nevertheless, the inclusion of spoken in addition to written English promises insight into potential variation.

For the multivariate analysis, Neumann and Evert (2021) extracted counts for 41 lexico-grammatical features from the texts in the data set. The counts were used as normalised, standardised and log-transformed frequencies (see above). After removing texts too short to reflect a reasonable distribution of linguistic features, their final set of texts amounted to 2,844 texts. In this paper, the focus is on the first two LDA dimensions, i.e. the most discriminatory ones reported by Neumann & Evert (2021), which will be explained in more detail in the next section.

[2] All scripts and the actual data set are accessible in the paper's online supplement at https://stephanie-evert.de/PUB/NeumannEvert2021/ .

4. Making Sense of Instantial Variation

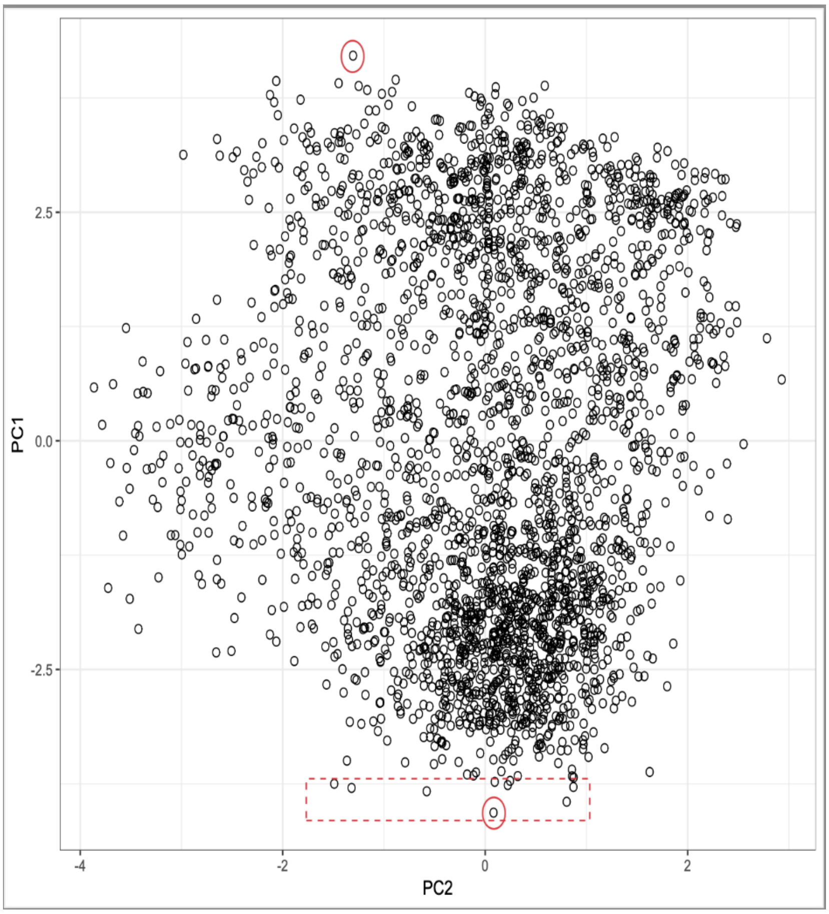
How can we link this quantitative analysis to systemic functional conceptualisation? Martin and Matthiessen (1991) observe that a purely typological perspective that categorises items as specimens of one of a set of types does not adequately capture the multiple criteria for which items need to be classified. They therefore propose a complementary multidimensional perspective on categories that captures similarities between items on one dimension and their differences on others. For this complementary perspective, they adopt the mathematical notion of topology introduced by Jay Lemke. Topology is characterised as "[turning] a set of objects into a *space* defined by the relations of those objects" (Lemke n.d. cited in Martin & Matthiessen 1991, 370). In this spatial representation, similar items are closer together and less similar items are further apart. Multiple criteria may lead to two items being closer along one dimension (e.g. vertical dimension), but further apart along another (e.g. horizontal dimension). This topological perspective is further discussed by Matthiessen (1995, 1875), describing classes as being mapped onto regions in continuous space. He illustrates this conceptualisation with a prototype-like representation of process types with a recognisable core and periphery. The topological perspective by Martin & Matthiessen (1991) and Matthiessen (1995) concentrates on the conceptual question of how categories for the analysis of language use (such as process types) can be represented in a way that does justice to the complexity of language.

The concept of topology is also at the heart of GMA, applied here to the adequate representation of the multidimensional characterisation of instances of language use in specific contexts. As mentioned above, GMA draws on the geometric distances between data points to interpret the configuration of texts identified by the multivariate analysis as near or distant on one dimension. It thus exploits the same mathematical concept. The exact position of each data point along a dimension is the result of the algorithm's weighting of all feature frequencies. This weight is computed such that those data points that share linguistic similarities in the form of similar feature frequencies are arranged near each other. The exact position of a data point along a dimension is then the result of summing all weighted feature frequencies.

PCA yields as many dimensions as there are feature variables (or data points, whichever is smaller), where each dimension reflects one specific constellation of shared feature frequencies out of the multivariate relationships, represented as distances of data points along this dimension. Conveniently, the PCA dimensions are organised systematically in descending order of variance between data points they describe. The additional amount of described variance drops considerably after a certain number of dimensions, as the higher dimensions will only capture minimal patterns in very few data points. This is conventionally taken as the cut-off point for deciding how many dimensions to consider.

GMA's topological perspective can be illustrated with the first two dimensions of Neumann and Evert's PCA. In Figure 2, each symbol represents one text characterised by the frequencies of all lexico-grammatical features included in the analysis (i.e. a feature vector). The two texts highlighted by circles are maximally distant along the vertical axis, which represents the first dimension. In line with the above reasoning, they are interpreted as also being linguistically maximally different along this dimension. In more concrete terms, they will be characterised by largely diverging feature frequencies. However, along this same vertical axis, the texts in the dotted box are quite near, but not so along the second, horizontal dimension. Linguistically, the texts in the box are interpreted as sharing similarities captured by the first dimension, but, at the same time, displaying differences captured by the second dimension. The interpretation of multidimensional vector space in GMA thus rests on interpreting Euclidian distances as linguistic differences. Exploring these distances also helps us assess variation between individual textual instances in their contribution to the overall clustering of texts. LDA is conceptually related to PCA in that it also captures differences between data points as distances in vector space, even if the exact computation works differently. Consequently, the spatial arrangement of data points can also be interpreted in similar ways.

Figure 2. First two PCA dimensions

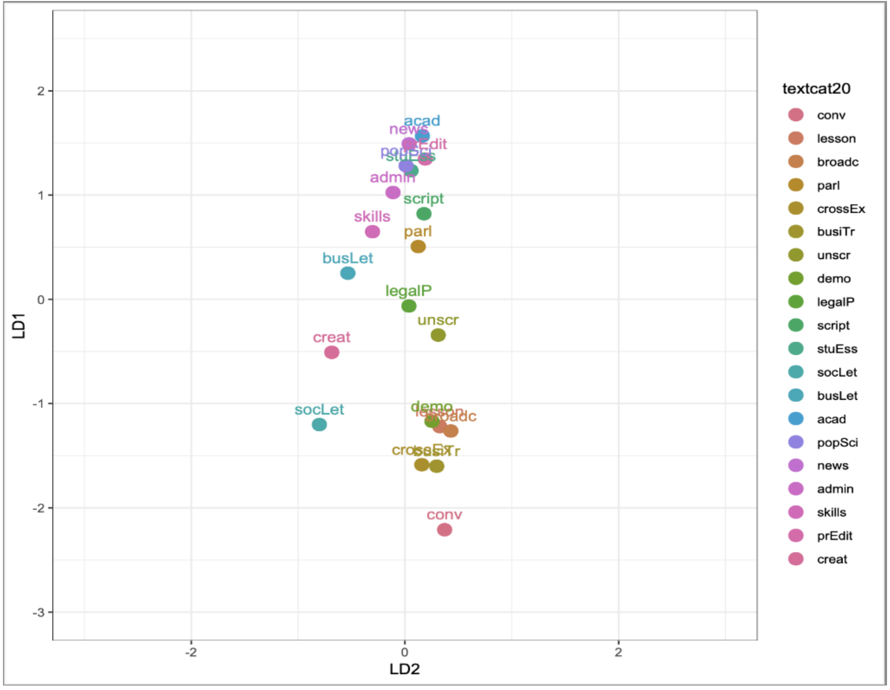


(*Adapted from Neumann & Evert 2021*)

Based on the combined exploration of the visualisation and the feature distribution along each dimension, Neumann and Evert (2021) interpreted the first LDA dimension as capturing variation between conceptual writing (positive pole) and conceptual speaking (negative pole), drawing on a conceptualisation by Koch and Oesterreicher (1985) which is related to Hasan's cross-classification between channel and medium (Halliday & Hasan 1989, 58). Towards the positive pole of the dimension, we find texts from prototypically written text categories such as academic writing and press editorials, but also scripted speeches. On the negative side we find not only spoken texts from conversations, classroom lessons and various broadcast interactions, but also (written) social letters. Social letters, in turn, mark the extreme category on the negative side along dimension 2. Since this dimension primarily brings out the dialogic character of these written texts as opposed to all other texts which are neutral along the dimension, Neumann and Evert label the dimension "dialogic written versus neutral".

Before exploring a full scatter plot, let us first inspect the picture that emerges when representing aggregated values for the two LDA dimensions. When aggregating the scores of texts along dimensions of the multivariate analysis, we get an overview of central tendencies for a given explanatory variable. The dots in Figure 3 represent so-called centroids, the combined mean values of each text category for the first two dimensions. They thus illustrate the mean-based perspective usually adopted by multivariate studies of register variation (see above).

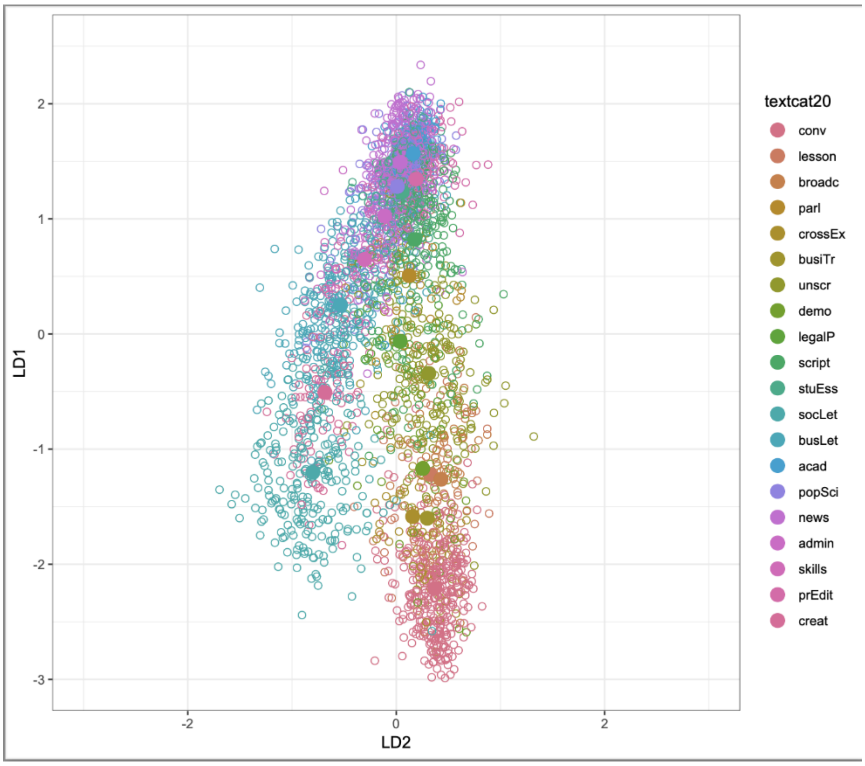
Figure 3. The first two LDA dimensions represented as centroids by aggregating data points



(*Adapted from Neumann & Evert 2021*)

Figure 3 paints a very clear picture of the 20 text categories in the data set.[3] A few dots overlap, but, generally, the categories are neatly separated. Variation across contexts of use appears to be categorically different. However, when adding all individual data points from which these centroids are aggregated, we get the scatter plot in Figure 4.

Figure 4. Scatter plot of the first two LDA dimensions representing all data points and centroids



(Adapted from Neumann & Evert 2021)

In Figure 4, we can discern distinctive clusters by data points in the same colour, especially at the lower end of the first, vertical,

[3] Since the fit between the notional text categories of the ICE Corpus and the actual registers within the respective communities is unclear, I will stick to the term ‘text category’ here. As an exploratory procedure, GMA can be used to discover potential new registers based on their shared linguistic feature distributions. However, these would then still need to be reviewed for shared contextual features.

dimension (LD1), where we can identify a cluster of red texts labelled as belonging to the category of casual conversation. Compared to the plot in Figure 3, the scatter plot also captures ubiquitous overlaps between the categories, suggesting that they are best conceived of as continuities. Such overlaps could be merely artefacts of the data, in cases where a text is misclassified in the category system of the corpus.[4] This issue cannot be ruled out in the present, heavily inductive approach. However, the combination of three different corpus components, compiled and labelled by three independent teams, mitigates the impact. The fact that most of the texts cluster in the same region of vector space across the three varieties gives credibility to the categorisation.

The centroids are still included in this visualisation and reveal one area that particularly illustrates the problems of aggregation: the three most extreme categories on the negative side of dimension 2 (horizontal axis), namely social letters (socLet), creative writing (creat) and business letters (busLet). While the centroids for these categories in Figure 3 are among the most clear-cut categories, Figure 4 shows that there are data points in a different colour right next to the centroids. These data points are categorised as belonging to a different text category than the centroid suggests. As mentioned above, the negative side of LD2 (the second dimension of the LDA) was interpreted as dialogic written. All three categories are located on this negative side of LD2. Visual inspection suggests that texts in these categories form a clear continuum. Social letters are most extreme and most likely to reflect conceptual speaking (Koch & Oesterreicher 1985). Business letters are also located on the negative side and are likely to contain more features typical of writing, while still also capturing some dialogic features. Creative writing appears to be a mixture of both, which can be explained by the combination of character dialogue and narrative parts in narratives. Generally, these texts are

[4] This is not a purely hypothetical issue in the International Corpus of English, as the general corpus design requires representing the given set of text categories across varieties of English. Given the expected cultural differences between English-speaking communities covered by the corpus, it is possible that a certain situational context supposed to be captured by a text category does not exist in a community. As a result, corpus compilers may force certain texts into a category despite its misfit.

characterised by higher frequencies of personal pronouns and interrogatives (with the exception of business letters) than the other texts in the data set. The overlap between these text categories therefore does not come as a surprise, despite the fact that their overall contextual characterisation will be quite different.

The visualisation not only offers evidence for widespread overlap between text categories, reflecting their 'fuzziness'. It also illustrates the massive instantial variation within each category. This is particularly interesting given the fact that LDA maximises within-group similarities between data points.

The variation between instances appears to be related to the paradigmatic organisation of language, which forces language users to make individual decisions at each choice point. The visualisation reveals that these individual decisions tend to be similar for given text categories (more so than for varieties of English, as Neumann and Evert 2021 note), but they also reveal individual preferences, which can be explained in part by the language users' individual semiotic biography (Taverniers 2021), that is, their current experience with language given their social background. However, despite a large amount of instantial variation, the visualisation suggests that the variation across texts is not open-ended. This would imply that each instance of language use would involve unpredictable choices at each choice point in reaction to the unique constellation of the dynamic unfolding of a particular situation, resulting in an idiosyncratic distribution of linguistic features. If this was the case, it would be impossible for the statistical technique to identify any clusters in the data. Arguably, some of the individual texts located in the areas of overlap between clusters will represent a change in the unfolding situational constellation as suggested by O'Donnell (2021), resulting in a blend of different text categories. However, the multivariate analysis indicates that this is not the default case.

Linguistic conceptualisations of variation as fuzzy sets, prototypes, or exemplars tend to focus on individual semantic or grammatical categories. This becomes clear in how Matthiessen (1995, 1874) refers to fuzzy classes: "Systemic theory has always held the view that options (classes) may be indeterminate or fuzzy." Therefore, he continues, "the terms of a system may represent regions on a cline rather than clear-cut classes." This paper argues that such conceptualisations can be transferred to complex entities such as texts made up of ensembles of individual features and to sets of texts. Or rather, in a usage-based perspective, we might conceive of the variation in frequencies of use across texts as the source of the fuzziness which is then observed in text categories. In this sense, the visualisation of texts in vector space offers empirical support of long-standing theoretical considerations in SFL and beyond. Specifically, it corroborates assumptions about widespread and systematic variation by situational context, and further specifies these by foregrounding instantial variation.

The understandable focus in much usage-based linguistic theorising is on demonstrating the importance of major types of variation, that is, on highlighting differences between clusters of texts. This focus, however, may be responsible for the neglect of individual variation. One question that appears particularly pertinent in this context is how big the effect of individual variation within the overall patterning of language is. To this end, new types of very large corpora are required that not only include texts by different language users, but also several texts across different situational contexts by the same language users.

Conclusion

This paper set out to examine predictions of patterned variation in systemic functional theorising by exploring quantitative empirical evidence of clusters of texts based on their distribution of linguistic features. Recent advances in visualising the results of sophisticated multivariate analysis of large corpora help in assessing the accuracy of these predictions. The visualisations provided support for claims about linguistic variation based on situational context. In line with systemic functional predictions, patterns do not form clear-cut categories, but seem to be characterised by fuzzy boundaries. Moreover, the visualisations also revealed widespread instantial variation. This variation does not contradict the theoretical assumptions, but it foregrounds a phenomenon that has not received much attention in SFL.

The amount of systematic variation across instances cannot be gauged through the small number of texts necessary in qualitative analysis. It is simply impossible to decide with the naked eye whether the difference in counts of a feature in individual texts is indicative of an overall pattern, as it cannot be known how these texts relate to the overall amount of variation in a given text category. As valuable as detailed qualitative analyses are for making sense of language, this specific question can only be decided on the basis of a systematic analysis of large enough samples of texts. The individual instance might appear inconspicuous, but through large-scale analysis, its contribution to the overall shape of language is revealed, thus reflecting its power to reinforce or change norms in a language. To this end, suitable corpora are required that allow for assessing the influence of idiosyncratic choices by the individual language user against the background of social and situational factors.

Acknowledgements

I am indebted to Stephanie Evert for the scripts that made the analyses in this paper possible and for numerous conceptual and methodological ideas and discussions about GMA. I gratefully acknowledge the two reviewers' very helpful comments, which helped me clarify the paper, as did the thoughtful proofreading by Anne-Coleman Webre.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*(s1), 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*, 3–43.

Biber, D. (1995). *Dimensions of register variation*. Cambridge University Press.

Biber, D., Egbert, J., Keller, D., & Wizner, S. (2021). Extending text-linguistic studies of register variation to a continuous situational space: Case studies from the web and natural conversation. In E. Seoane & D. Biber (Eds.), *Studies in corpus linguistics* (Vol. 103, pp. 19–50). John Benjamins Publishing Company. https://doi.org/10.1075/scl.103.02bib

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. MIT Press.

Brezina, V., & Meyerhoff, M. (2014). Significant or random?: A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, *19*(1), Article 1. https://doi.org/10.1075/ijcl.19.1.01bre

Diwersy, S., Evert, S., & Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In B. Szmrecsanyi & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (pp. 174–204). de Gruyter.

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics Conference 2011, University of Birmingham, UK, 20-22 July 2011*. Corpus Linguistics Conference 2011, Birmingham, UK. http://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies. New theoretical and methodological traditions* (pp. 47–80). de Gruyter.

Grafmiller, J., Szmrecsanyi, B., Röthlisberger, M., & Heller, B. (2018). General introduction: A comparative perspective on probabilistic variation in grammar. *Glossa: A Journal of General Linguistics*, *3*(1). https://doi.org/10.5334/gjgl.690

Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The International Corpus of English*. Clarendon Press.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, *17*(3), 241–292.

Halliday, M. A. K. (1999). The notion of 'Context' in language education. In M. Ghadessy (Ed.), *Text and context in functional linguistics* (pp. 1–24). John Benjamins Publishing Company.

Halliday, M. A. K. (2013). Meaning as choice. In L. Fontaine, T. Bartlett, & G. O'Grady (Eds.), *Systemic functional linguistics: Exploring choice* (pp. 15–36). Cambridge University Press.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford University Press.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4th ed.). Routledge.

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In *Romanistisches Jahrbuch* (Vol. 36, pp. 15–43). de Gruyter.

Kortmann, B., Lunkenheimer, K., & Ehret, K. (Eds.). (2020). *eWAVE - The Electronic World Atlas of Varieties of English*. Zenodo. 10.5281/zenodo.3712132

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide: A Journal of Varieties of English*, *37*(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Martin, J. R., & Matthiessen, C. M. I. M. (1991). Systemic typology and topology. In F. Christie (Ed.), *Literacy in social processes: Papers from the Inaugural Australian Systemic Functional Linguistics Conference, Deakin University, January 1990* (pp. 345–383). Centre for Studies of Language in Education, Northern Territory University.

Matthiessen, C. M. I. M. (1993). Register in the round: Diversity in a unified theory of register analysis. In M. Ghadessy (Ed.), *Register analysis: Theory and practice* (pp. 221–292). Pinter.

Matthiessen, C. M. I. M. (1995). Fuzziness construed in language: A linguistic perspective. *Proceedings of 1995 IEEE International Conference on Fuzzy Systems.*, *4*, 1871–1878. https://doi.org/10.1109/FUZZY.1995.409935

Neumann, S. (2014). *Contrastive register variation: A quantitative approach to the comparison of English and German*. de Gruyter Mouton. https://doi.org/10.1515/9783110238594

Neumann, S., & Evert, S. (2021). A register variation perspective on varieties of English. In E. Seoane & D. Biber (Eds.), *Corpus-based approaches to register variation* (pp. 143–178). John Benjamins Publishing Company.

O'Donnell, M. (2021). Dynamic modelling of context: Field, Tenor and Mode revisited. *Lingua*, *261*, 102952. https://doi.org/10.1016/j.lingua.2020.102952

Taverniers, M. (2021). Modelling interfaces with context in SFL: Stratification, instantiation, metafunctions. *Functions of Language*, *28*(3), 291–314. https://doi.org/10.1075/fol.20015.tav

Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, *28*(4), 421–450. https://doi.org/10.1111/j.1467-971X.2009.01606.x

About the author

Stella Neumann is a full professor of English Linguistics at RWTH Aachen University, Germany. Her research is mainly concerned with understanding how linguistic variation shapes language. Grounded in systemic functional linguistics, it draws on quantitative corpus and experimental methods to study variation across registers, varieties, languages and in translation. As a corollary of this interest, her work also has a methodological component.

بحوث جامعية

دورية تصدر عن كلية الآداب والعلوم الانسانية بصفاقس

ر.د.م.م – ISSN – 2811-6585

Recherches Universitaires

كلية الآداب والعلوم الانسانية بصفاقس

صندوق بريد 1168، صفاقس 3000 تونس

الهاتف: 557 670 74 (216) - 558 670 74 (216)

الفاكس: 540 670 74 (216)

الموقع الإلكتروني: www.flshs.rnu.tn

العدد 19 خاص في جز أين باللغة الإنجليزية

أشرف على اعداده مخبر البحث في مقاربات الخطاب

الجزء الأول

(أبريل 2025)

Academic Research

N°19, Vol.1, Special Issue

Published by the **Laboratory on Approaches to Discourse** (LAD/LR13ES15)
and the **Systemic Functional Linguistics Association of Tunisia** (SYFLAT*)*