Extending the Scope of Lectometry
Zur Erweiterung des Anwendungsbereichs der Lektometrie

Abstract: This special issue brings together a collection of papers that highlight recent advances in the field of lectometry. Lectometric research measures distances between various types of language varieties, such as dialects or sociolects. In addition to briefly sketching the development of the field, this introductory paper surveys the theoretical, methodological, and practical contributions a lectometric approach to language has to offer. We furthermore identify blind spots within the current scope of lectometric work and lay out suggestions for future avenues the field could go down. The paper concludes by introducing the various contributions to this special issue of “Zeitschrift für Dialektologie und Linguistik” and by outlining how each of those papers contributes to extending the scope of lectometry.

Keywords: lectometry, dialectometry, sociolectometry, language variation, register variation, quantitative methods in linguistics


Schlagwörter: Lektometrie, Dialektometrie, Soziolektometrie, Sprachvariation, Registervariation, quantitative Methoden in der Linguistik
1. Defining lectometry

The central aim of lectometric research is to measure distances between lects, where lects are defined as language varieties that differ along any of various dimensions, such as dialects and regiolects, which vary along the geographical axis, sociolects, which vary according to social characteristics of language users, or lects representing different styles. To meet this goal, lectometry typically takes an aggregate perspective, measuring distances between lects in a more comprehensive way rather than on the basis of single linguistic variables. While the use of the term lectometry is a more recent evolution (for example Ruette/Speelman/Geeraerts 2011), the field is related to other research traditions in linguistics that take an aggregate approach. Dialectometry, for instance, which evolved out of the need to objectively measure distances between what are traditionally considered the base dialects of a language, has been around since the 1970s (Séguy 1971). This field of study has been gaining ground over the last decades, most notably in the Salzburg (Goebel 1984) and Groningen school (Nerbonne/Heerenga/Kleiweg 1999). Similarly, studies in register analysis have been using quantitative techniques to examine the similarities and differences between texts from different registers on an aggregate level since the late 1980s (Biber 1988). A more recent tradition is sociolectometry which examines differences between language varieties on the basis of social features (Geeraerts/Grondelaers/Speelman 1999). The field of lectometry then can be considered an overarching research program that brings these early traditions and novel approaches together.

2. Why pursue a lectometric approach to language?

So what is there to gain for linguists in conducting lectometric studies? This question begs a theoretical, a methodological, as well as a practical response. On the theoretical level, the usage-based nature of modern approaches to linguistics implies that linguistic research needs to rely on empirical observation (for example Bybee/Hopper 2001, Dąbrowska/Divjak 2015, Kemmer/Barlow 2000, Tomasello 2001). Recently, scholars have argued that this entails coming to terms with variation between lects of any kind: linguistic variation is always multifactorial (Geeraerts 2005, Croft 2009, Dąbrowska 2015a, Harder 2003, Majid/Burenhult 2014, Schmid 2016). The field of lectometry hinges on this view, as it explicitly examines how the multifactorial nature of language influences the relationship between different lects. Furthermore, recent studies in linguistics have shown that the loss of traditional dialects has led to increased regiolectal variation in many European languages (Auer 2005). This shift opens up new questions regarding the relationship between different varieties of a single language (for example the relationship between Colloquial Belgian Dutch in Flanders and Standard Dutch as in Geeraerts/Grondelaers/Speelman 1999 or between European and Brazilian Portuguese as in Soares da Silva 2010). In this context, diachronic questions also rise regarding processes of convergence and divergence between
lects, increasing or decreasing variability, and informalisation vs. standardisation (cf. Geeraerts 2018). Lectometric studies are, by their nature, ideally suited to examine questions like these.

On the methodological level, the availability of larger datasets and the use of quantitative methods has opened up new ways of examining linguistic variation. In contrast to, for instance, dialectometric studies, which are traditionally based on dialect atlases, or to variationist sociolinguistic studies, which often rely on sociolinguistic interviews, lectometric research methods can also be applied to attitudinal and perceptual data or to (synchronic and diachronic) corpus material. Using lectometric techniques can then offer novel insights into variation phenomena in a wide range of naturalistic language data. We will discuss some of these methodological advances in more detail below.

In addition to theoretical and methodological gains, lectometry also has practical advantages to offer. As explained above, using lectometric approaches to study language variation and change helps linguists to develop a better understanding of the actual state of a language or lect and its relation to other languages or lects. It also offers insights in the ongoing dynamics of languages and lects. Such knowledge is vital for well-informed language policy making. In that sense, a lectometric approach to linguistics can serve as one of the cornerstone of evidence-based language policies and have an impact on the way language is approached in society at large.

3. Recent expanses in the field of lectometry

In the past few years, the field of lectometry has witnessed a substantial expansion. We identify four axes along which new ground has been explored by recent lectometric work. Firstly, the types of lects considered in lectometric studies have increased significantly. While dialectometry traditionally limited itself to geographic variation, more recent work has applied the lectometric approach to study a more varied range of sociolects, as well as stylistic differences for an ever-growing set of languages around the world. Examples for sociolectometric research are Plevoets (2008) and Ruette/Speelman/Geeraerts (2011) who work on Dutch. For work on stylistic variation we can cite Berber Sardinha/Kauffmann/Mayer Acunzo (2014) on Brazilian Portuguese or Asención-Delaney/Collette (2011) on L2 Spanish. When we look at the expansion in registers and text types that are studied in lectometry we can for instance cite work by Diwersy/Evert/Neumann (2014) and De Sutter/Delaere/Plevoets (2012) who compare translated and non-translated texts from an aggregate perspective or Karsdorp/Fonteyn (2019) who focus on styles in folk tales. This expansion of the type of lects studied with lectometric approaches is also clearly represented in this special issue. While Ghyselen/Plevoets/De Sutter (this issue) chart stylistic variation in Belgian Dutch, they pay attention to the lects of younger and older speakers and even discuss individual variation between the speakers in their corpus. This attention to sociostylistic variation is also present in the contribution of Hilte/Vandekerckhove/Daelemans (this issue) where the language of
teenagers of different educational levels, ages, and genders is compared. Furthermore, **HILTE/VANDEKERCKHOVE/DAELEMANS** (this issue) also focuses on digilects, a novel type of lect based on online chat materials. This is also the case in the contribution by **BOHMANN** (this issue) who compares these digilects to more traditional genres. The paper by **PLEVOETS** (this issue) then, follows up on the recent trend in lectometric research to focus on the characteristics of translated versus non-translated texts. The fact that lectometry is nowadays used to study lects in a wide variety of languages around the world is also clear from the contributions in this special issue. Although Dutch is strongly represented with three contributions (**GHYSLENS/PLEVOETS/SPPELMAN** this issue, **HILTE/VANDEKERCKHOVE/DAELEMANS** this issue and **PLEVOETS** this issue), **BOHMANN** (this issue) focuses on world Englishes, **SOUSA/DUBERT** (this issue) study Galician dialects in Spain and the work by **AURREKOETXEA** et al. (this issue) widens the scope to lectal variation in a non-Indo-European language, namely Basque.

Not only the types of lects and the variety of languages dealt with in lectometry have expanded, the field is increasingly concerned with a wider selection of linguistic features. Over the past few decades, lectometric work has come to cover lexical variation (for example **GEERAERTS/GRONDELERS/SPPELMAN** 1999, **RUETTE/EHRET/SZMRCSANYI** 2016, **WIELING/UPTON/THOMPSON** 2014, **FRANCO/GEERAERTS/SPPELMAN** 2019), morphosyntactic variation (for example **SZMRCSANYI** 2014), as well as phonetic, and phonological variation (for example **MONTEMAGNI** et al. 2013). What is more, recent work strives to include linguistic features of various levels of linguistic structure to arrive at a better representation of the variation exhibited in the lects under study (for example **HEERINGA/HINSKENS** 2014 who take into account phonetic, lexical as well as morphological features or **SPRUIT/HEERINGA/NERBONNE** 2009 who include pronunciation, lexical, and syntactic features). This expansion of the type of linguistic features considered in lectometric work can also clearly be witnessed in this issue. **GHYSLENS/PLEVOETS/SPPELMAN** (this issue) consider both phonetic/phonological and morphosyntactic features, just like **AURREKOETXEA** et al. (this issue) who additionally include lexical features. This last type of features is also considered in the contribution by **HILTE/VANDEKERCKHOVE/DAELEMANS** (this issue) who combine them with typographic variation and even emojis, and in the paper by **BOHMANN** (this issue) who studies them alongside grammatical and discourse features.

Thirdly, the type of data used to conduct lectometric analyses has been evolving. Traditionally, dialectometric analyses were based on atlases or dictionary materials collected through dialectological questionnaires (cf. **GOEBL** 2010: 433–436 for an overview). This type of study is of course still relevant today, especially as recent technological advances make the availability and the analysis of large-scale dialectological datasets increasingly more feasible (**NERBONNE/KRETZSCHMAR** 2003). However, as lectometric studies have expanded their scope from dialects to include other types of lects, they also increasingly rely on naturalistic, non-elicited linguistic material. A prime example is **SZMRCSANYI** (2013), which is described as “a study in corpus-based dialectometry”. **SZMRCSANYI** analyses a large number of variable linguistic features in FRED, the “Freiburg Corpus of English Dialects”, to investigate the aggregate lectometric land-
scape of the UK. Another example is Grieve (2016) who uses data from regional newspapers to investigate the patterning of a large amount of grammatical alternations in American English. Additionally, many lectometric studies use corpora that are in some way tailored to a specific goal. For example, Corrigan/Mearns/Moisl (2014) have used historical data to examine variation within a lect across time. Geeraerts/Grondeelaers/Speelman (1999) compiled corpora of clothing and football terms to examine convergence and divergence between Belgian and Netherlandic Dutch. In line with this trend, the contributions in this special issue cover a range of different data sources: dialectological atlas data (Aurrekoetxea et al. this issue, Sousa/Dubert this issue), translation corpora (Plevoets this issue), data from social media (Bohmann this issue, Hilte/Vandeckerckhove/Daelemans this issue) and a combination of different types of sociolinguistic data (Ghyselen/Plevoets/Speelman this issue).

Finally, this expansion of lectometric studies to new types of data has also brought about the need for novel analysis techniques. Dialectometric analyses are increasingly relying on quantitative methods tailored to spatially structured data to examine the distance between the dialects in older dialectological materials (for example De Vriend/Swanenberg/Van Hout 2011, Meschenmoser/Pröll 2012, Wieling/Nerbonne/Baayen 2011) and in more recently collected regionally stratified data (for example De Pascale 2019, Grieve/Speelman/Geeraerts 2011, Plevoets 2008, Wieling/Upton/Thompson 2014). While traditional methods for stylistic lectometric analyses are still very popular (see for instance the papers collected in Berber Sardinha/Pinto 2014), the field of stylometry has recently also been incorporating methods like machine-learning algorithms and NLP tools (cf. Neal et al. 2017). The papers in this special issue reflect this wave of methodological innovations in lectometry. Plevoets (this issue) introduces correspondence regression as a novel methodological tool for lectometry. Aurrekoetxea et al. (this issue) describe a method that can be used to obtain the minimal number of linguistic features that is necessary to distinguish between lects.

4. Challenges ahead

Despite these recent advances in lectometry, certain aspects remain underexplored in the field. We identify three desiderata for the future. First, lectometric research on perception and attitudinal data is still largely lacking (with notable exceptions like Impe/Geeraerts/Speelman 2009). The second and third aspect that should be a focal point in future work concerns communication and collaboration both within the field of lectometry and with other fields outside lectometry that are also concerned with language variation from an aggregate perspective. Different strands of lectometric research (for example dialectometry, register studies in the tradition of Biber, sociolectometric work) do already interact and cross-pollinate (cf. Wieling/Nerbonne/Baayen 2011 who combine socio- and dialectometry; Rquette/Speelman/Geeraerts 2011 who marry a focus on register with one on sociolects). Yet, this mutual fertilization should
be intensified as it fuels advances in the field. Likewise, attempts have been made to bring together lectometric research and studies in fields like typology that take a similar quantitative and aggregating approach to measuring distances between languages (for example Wälchli/Szmrecsanyi 2014), but much room is left for improved communication and collaboration.

5. Contributions to this special issue

We conclude this introduction by introducing the contributions to this special issue in more detail.

Offering a new methodological approach, Koen Plevoets’ contribution is an exercise in the description and application of correlation regression, a novel statistical method, on previously analysed data (translated texts from the Dutch Parallel Corpus and from a corpus of subtitles). The proposed method is similar to multidimensional scaling analyses (MDS) in that it tries to uncover underlying dimensions or latent variables that can account for the association/correlation between response and predictor variables. Plevoets suggests that while a lectometric approach can provide insights into the distance between varieties, what drives these distances (for example a certain combination of registers, lexical items, etc.) has not been explored so far. Plevoets tackles this open question by first outlining the mathematical model needed to measure the impact of independent predictors (for example choice of language, register) on the frequency of a (multinomial or binary) response variable. In traditional lectometric analyses, this response variable is used as one of the many linguistic predictors rather than as a response variable. The paper then introduces correspondence regression – see also Ghyselen/Plevoets/Speelman (this issue) – and uses it to analyse the correspondence between multiple predictor variables and a multinomial response variable. The association between the variables is considered indicative of latent variables which can be visualised with association graphs. The paper showcases the use of the method in two case studies that compare the frequency of 13 lexical variables in translated linguistic material (Dutch translated from French/English) to their frequency in the original language (non-translated Dutch). The first of these studies highlights that register differences are dependent on the source language – refuting Baker’s (1993) claim about the universality of translational language across registers. With regard to the underlying, latent, variables, the first study identifies (non-)standardness and formality as latent variables for written translations. In other words, the differentiation between non-standard and standard variants and between informal and formal variants accounts the most for the distribution of the predictor and response categories. The second case study identifies (non-)standardness as a latent variable in audiovisual translations while the second latent variable seems to be related to compactness.

Similar to Plevoets’ contribution, Lisa Hilte, Reinhild Vandekerckhove and Walter Daelemans’ study focuses on the use of non-standard features in Belgian Dutch, yet they zoom in on variation in computer-mediated communication (CMC).
While Plevoets identified non-standardness and formality as the defining criteria underlying linguistic variation in written translations in a bottom-up fashion, Hilte/Vandekerckhove/Daelemans’ (this issue) contribution specifies non-standard and regional features a priori and supplements the analysis with socio-demographic data. Besides its focus on non-standard regional variants, the study also includes linguistic features that are specific to the online writing culture (so-called “new” vernacular forms) in its analyses. The data stems from social media posts (Whatsapp and Facebook messages) by adolescents aged 13 to 20 and includes 11 different variables ranging from typographic expressive markers (for example emojis), the use of oral, informal speech (for example English lexemes), to compressed words or utterances (for example abbreviations and acronyms). Essentially, Hilte/Vandekerckhove/Daelemans (this issue) analyse the correlation between these different types of non-standard writing and adolescents’ socio-demographic background (gender, age, education) using generalised linear mixed-effects models. The results of their study show that all sociolinguistic variables impact the degree of non-standardness, with a higher use of non-standard variants by younger speakers and a significant increase of expressive and oral non-standard variants in female speech. These results confirm earlier sociolinguistic work that observed that female adolescents seem to prefer non-standard use compared to female adults and that female speakers seem to shift at some point to the more standard forms. At the same time, boys do not seem to shift in their usage frequency between adolescence and adulthood. Gender-related differences also become apparent in the type of non-standard variants used with girls preferring expressive markers and boys preferring oral features. The authors conclude by suggesting that boys and girls do not share the same digilect. While the methods employed in Hilte/Vandekerckhove/Daelemans (this issue) may be considered as more established in research on linguistic variation, their application to CMC is innovative.

Adding a layer to these larger sociodemographic groupings, Anne-Sophie Ghyselen, Koen Plevoets and Dirk Speelman’s study zooms in on individual- as well as community-level grammars by analysing the linguistic repertoire of ten highly educated Dutch speaking women in the Ghent area (Belgium). The authors use three complementary multivariate statistical techniques, namely correspondence regression (see also the contribution by Plevoets), multidimensional scaling, and cluster analysis, to explore covariance patterns of linguistic variants on the level of the community and the individual. The data comprise 11 phonological and 14 morphosyntactic variables that are characteristic of one or more varieties of Dutch: the traditional Ghent dialect, Standard Dutch or varieties in between these two. Similar to Hilte/Vandekerckhove/Daelemans’ (this issue) study, then, Ghyselen/Plevoets/Speelman’s (this issue) contribution focuses on the use of regional and standard-like variants and the correlation between the frequency of these variants with sociodemographic factors, namely age, speech setting, and speaker. In contrast to Hilte/Vandekerckhove/Daelemans (this issue), who take age, gender, and education as the basis for a speaker’s sociodemographic background, Ghyselen/Plevoets/Speelman (this issue) focus specifically on the situational context, i.e. speech setting, and on intra- and inter-
ter-individual variation. Aggregated across all speech settings, results indicate that variants cluster according to “localness” or standardness (see Plevoets’ contribution) in that non-standard, local variants cluster against standard, exogenous variants (i.e. variants of other regional dialects). Results also point to inter- and intra-individual variation: whereas on the community level the ten speakers show homogeneity in their use of standard forms in a standard language test, they display more idiosyncratic behaviour in their use of dialectal variants with some speakers exhibiting a more limited linguistic repertoire than others (i.e. monoglossic behaviour) and with a difference in dialect competence between older and younger speakers. The combination of the three statistical techniques showcases the usefulness of applying multiple types of analysis to arrive at more reliable results. As in Plevoets’ contribution, the authors identify the underlying dimension of linguistic variation and aim to align this dimension with one of Auer’s five preconceived types of community repertoires. What is more, their focus on intra-individual and inter-individual variation shows how lectometric approaches can contribute to recent discussions in various fields of linguistics regarding the correlation between individual- and community-level grammars and the importance of teasing the two apart (see for instance Petré/Van de Velde 2018 for historical linguistics, Tamminga/Mackenzie/Embick 2016 for variationist sociolinguistics, Dąbrowska 2015a, Dąbrowska 2015b for cognitive linguistics).

While the previous three contributions all focus on Dutch and varieties thereof, the study by Bohmann offers a lectometric analysis of social media data tapping into varieties of English. Complementing the picture offered by Hilte/Vandekerckhove/Daelemans (this issue) on the register of social media in Dutch Whatsapp and Facebook messages, Bohmann investigates the text type properties of Twitter discourse in comparison with traditional text types such as news reports and narrative fiction. Data stem from the author’s own collection of tweets and the available ten national corpora of the International Corpus of English (ICE) series. Bohmann’s contribution thus zooms out again to the aggregate level glossing over idiolectal and intra-individual variation (cf. Ghyselen/Plevoets/Speelman this issue) and attends to the hitherto unaddressed issue regarding the structural relationship between Twitter discourse and offline registers. Frequencies of 236 individual linguistic features were extracted from the corpus and subjected to factor analysis in order to determine the underlying dimensions or feature bundles (i.e. highly correlated features whose frequency distribution across texts is similar) and the degree to which individual texts align with these dimensions. The factor analysis identifies ten necessary dimensions or feature bundles, among them “colloquial markedness”, “involved vs informational production”, and “collaborative communicative orientation”, whose characteristics – that is, the contributing features and the most typical text types associated with them – are discussed in detail. Taking this multi-dimensional nature of register variation into account, Bohmann argues that Twitter discourse can be characterised not only by its strategically colloquial aspect that tries to represent oral conversation in a written medium but also by its collaborative and involved style. Bohmann’s approach is thus novel in that it is the first to provide a detailed account of the characteristics of language on Twitter while also cautioning the
Extending the Scope of Lectometry

reader against a simplified, binary relationship between Twitter and other offline registers. With its general description of the language on Twitter, Bohmann’s study adds to Hilte/Vandekerckhove/Daelemans’s (this issue) contribution which adopts a more social perspective on the use of social media in adolescents. Bohmann’s paper further complements Ghyselen/Plevoets/Speelman’s (this issue) study which also investigates the correlation between different text types or registers and the linguistic features analysed.

Moving on to another geographic area, namely the Iberian Peninsula, the last two contributions focus on Spanish loanwords in Galician and on the Basque dialect area. The contribution by Sousa and Dubert-García combines a sociolinguistic interest in the spread of lexical forms between different languages in contact situations with a dialectometric approach to study linguistic variation. More specifically, the authors analyse the variation between Spanish loanwords and words in native Galician – a language which is historically related to neighbouring Portuguese – focusing on 68 lexical items sampled in the survey data of the “Atlas Lingüístico Galego” (1974–1976). Comparing and aggregating over the number of localities where Spanish loanwords were used vis-à-vis Galician native words, Sousa and Dubert-García’s study shows that geographic proximity to the Spanish-speaking areas (eastern Galicia), the existence of main communication routes, and/or low population density correlates with a high frequency of Spanish loanwords while geographic distant areas (northwestern Galicia) exhibit a low frequency of Spanish loanwords. Sousa and Dubert-García thus showcase how dialectometric tools can be applied to explore inter-systemic variation between languages rather than intra-systemic variation (as in Hilte/Vandekerckhove/Daelemans’ this issue and Ghyselen/Plevoets/Speelman this issue).

Finally, the more technical paper by Aurrekoetxea/Clua/Iglesias/Usubia-Ga/Salicrú (this issue) describes a novel method to determine the set of linguistic features that are minimally necessary to discriminate between dialect groups. Exemplifying their method with the Bourciez Corpus, which samples data from different localities in the Basque territory of Spain, they show that their bottom-up approach can identify the appropriate number of clusters as well as the most informative linguistic features. Aurrekoetxea/Clua/Iglesias/Usubia-Ga/Salicrú (this issue) method thus complements other, more traditional, dialectometric techniques used in the other studies in this issue and showcases how in cases where limited (economic but also linguistic) resources are available, conclusions drawn on a small sample of linguistic features can still be valid.

References


Laurarosseel / KarlienFranco / MelanieRöthlisberger


De Pascale, Stefano (2019): Token-based vector space models as semantic control in lexical sociolectometry. [PhD dissertation, KU Leuven].


Extending the Scope of Lectometry


MAJID, ASIFA/NICLAS BURENHULT (2014): Odors are expressible in language, as long as you speak the right language. In: Cognition 130 (2), 266–270.


LAURA ROSSEEL
Linguistics and Literary Studies (LIST), Faculty of Arts and Philosophy, Vrije Universiteit Brussel, Pleinlaan 11, 1050 Brussels, Belgium, e-mail: <Laura.Rosseel@vub.ac.be>

KARLIEN FRANCO
Department of Linguistics, Quantitative Lexicology and Variational Linguistics, Blijde-Inkomststraat 21 – box 3308, 3000 Leuven, Belgium, e-mail: <karlien.franco@kuleuven.be>

MELANIE RÖTHLISBERGER
Deutsches Seminar, Schönberggasse 2, 8001 Zürich, Switzerland, e-mail: <melanie.roethlisberger@es.uzh.ch>