Bernhard Wälchli (Stockholm University)
and Benedikt Szmrecsanyi (University of Leuven)

Introduction: The text-feature-aggregation pipeline in variation studies

1. Objectives of this volume

This volume aims to bridge sub-disciplinary boundaries in the study of linguistic variation – be it language-internal or cross-linguistic.[1] Even though dialectologists, register analysts, typologists, and quantitative linguists all deal with linguistic variation, there is astonishingly little interaction across these fields. However, the fourteen contributions in this volume show that dialectology, register analysis, typology, and quantitative linguistics share, in point of fact, many research goals and methodological concerns. In particular, the contributions to this volume converge with regard to all four of the following themes:

a) they seek to explore linguistic variation, within or across languages;
b) they are based on usage data, that is, on corpora of (more or less) authentic text or speech of different languages or language varieties;
c) they are concerned with the joint analysis (also sometimes known as "aggregation" or "data synthesis") of multiple phenomena, features, or measurements of some sort;
d) they marshal some sort of quantitative analysis technique to analyse the data;

In short, the volume deals with the quantitative analysis of linguistic variation from text via features to aggregation – what we take the liberty to call the *text-feature-aggregation pipeline in variation studies*. Why devote a volume to this pipeline? Succinctly put, there are two good reasons. First, themes a) through d) have in recent years become increasingly important in linguistics at large. Second, we firmly believe that dialectologists, register analysts, typologists, and quantitative linguists can, and should, draw mutual inspiration in many respects, thus overcoming petty sub-disciplinary narrowness.

Our point of departure is the fact that variation is increasingly seen as a "core explanandum" (Adger and Trousdale 2007: 274) in linguistics, and the

[1] The present volume builds on a workshop held at the Freiburg Institute for Advanced Studies (FRIAS) in February 2011 about "Cross-linguistic and language-internal variation in text and speech".

present volume aims, among other things, to illustrate how different disciplines and schools of thought in linguistics have different traditions of taking variation into account. Second, our focus on corpus studies is ultimately indebted to the usage-based turn in linguistic theory (e.g. Bybee 2010; Tomasello 2003), which posits that grammatical knowledge is experience-based, and should thus be studied by investigating language in use (*parole*), avoiding data reduction and/or abstraction as much as possible. Yet, our volume differs from many approaches in usage-based linguistics due to its dedicatedly cross-varietal focus. It does, however, share the view of usage-based approaches according to which language is contextually-embedded, the fact that utterances and their parts have particular situational meanings and always serve particular functions. In this sense, all utterances and their parts, words and constructions, are context-dependent and context-renewing, very much as has been argued for a long time in Conversation Analysis (Atkinson and Heritage 1984). Third, big-picture analyses are increasingly popular – not only in dialectology (e.g. Nerbonne 2008) or register analysis (e.g. Biber 1988), but also, to some extent, in cross-linguistic typology (e.g. Cysouw to appear). Lastly, quantitative methods allow the analyst to generalize over text-based data without claiming that language is a system with idealized speaker/hearers.

The present volume also continues the tradition of interface explorations spearheaded by the 2004 volume *Dialectology Meets Typology*, edited by Kortmann (2004). The difference is that the present volume has a more methodological orientation, and is more inclusive in that it also includes register-analytic work as well as work in the Quantitative Linguistics tradition. To be sure, this inclusiveness has been a challenge in assembling the chapters into a coherent volume. Dialectologists and cross-linguistic typologists have interestingly different notions of how a "joint analysis of phenomena, features, or measurements" should be designed. Different subdisciplines in linguistics also tend to tap into different sorts of corpora, as we shall see presently. But this – in our view – inevitable and, indeed, illuminating and enjoyable pinch of heterogeneity aside, the common denominator of the studies collected in this volume is that they explore variation on the empirical basis of "real" data, with an eye toward the big picture and utilizing quantitative analysis methods.

Aside from the relevance of these common issues, dialectologists, register analysts, typologists, and quantitative linguists can profit from each other in various respects.

First of all, we all use increasingly similar methods of analysis which can be subsumed under the heading of "aggregation". However, rather than talking about methods in abstract terms we think it is important to see the

methods applied to concrete data in order to understand their full potential and their limitations. You can learn a lot about the methods you use yourself by seeing your colleagues applying them to different data in slightly different ways and drawing conclusions slightly different from the ones you would have expected.

Second, all four linguistic sub-disciplines somehow aim at taking the full range of linguistic variability into account. However, the fact already that every sub-discipline does this in a very specific way shows that there are many different ways to address linguistic diversity and, as we will elaborate in Section 2, empirical attempts to account for linguistic variability go hand in hand with a need to keep some aspect of variation constant. Since our ultimate goal is a better understanding of linguistic variability in general, all of us should be interested in what other researchers, exploring other aspects of linguistic diversity, do.

Third, there are many ways in which the four sub-disciplines are not neatly distinct from each other. The typological study of genealogically related or closely related languages, sometimes called "intragenealogical typology", is in many respects intermediate between typology and dialectology (see the papers by von Waldenfels and Verkerk). Corpus-based dialectometry (see, e.g., Szmrecsanyi's contribution) profits a great deal from the methods of register analysis. Koptjevskaja-Tamm and Sahlgren show how lexical typology and register analysis can profit from each other. This volume thus invites the reader to discover the fuzziness of boundaries between dialectology, register analysis, typology and quantitative linguistics (along with some convergence of their methodologies).

The structure of this introduction is predetermined by the characteristics (a-d) listed above, neatly creating a processing pipeline that is – in some way or another – adhered to in most papers in this volume. Linguistic variation (Section 2) in usage data (Section 3) is captured in terms of features (Section 4) which are aggregated using some sort of quantitative analysis technique (Section 5). Rather than simply abstracting the papers in this volume we refract the discussion into these four colours of their spectrum.

2. Varieties and variation phenomena considered

All three disciplines dialectology, register analysis, and typology are traditionally engaged in classifying linguistic varieties into groups. Technically speaking, we could say that they all partition their objects of study into clusters. Dialectology groups dialects into dialect groups, register analysis groups texts into registers, and typology groups languages into types (Table 1).

Table 1. Partitioning of varieties in different approaches.

	Base unit	Cluster	Set of varieties considered
Dialectology	Dialect	Dialect group	All contiguous dialects of a language
Register analysis	Text	Register	A diverse corpus of texts reflecting a broad range of situational, social, and communicative task variation
Typology	Language	Type	A diverse sample of languages

All three approaches have in common that grouping is based on directly observable features of the varieties to be classified rather than on historical considerations or metadata. (This does not mean that the choice of data is independent of metadata, on the contrary, see below.) Typology contrasts with historical linguistics, which classifies languages into genealogical families and subfamilies (genera), and to areal linguistics, which searches for areal groupings of languages. Dialectology does not seek to reconstruct a genealogical tree of dialects, but rather to classify dialects according to their structural similarities. Register analysis contrasts with the study of genres. Metadata such as spoken vs. written or fiction, broadcast, interview, letter, conversation, planned speech etc. can be highly relevant for the interpretation of a resulting classification, but these do not constitute the criteria for classification.

Now if language varieties are grouped into clusters according to structural similarities, the question arises as to which structural similarities are taken into account. In the register analytic terms of Biber (1985) we can distinguish *microscopic* and *macroscopic variation*. Microscopic analysis provides an analysis of the functions of particular features in particular varieties. Macroscopic analysis attempts to define the overall dimensions of variation within a set of varieties. Since the 1980s register analysis has undergone a macroscopic shift, mainly due to the work of Douglas Biber and others. Dialectology has undergone a similar shift through Séguy's and Goebl's pioneering work on dialectometry (e.g., Séguy 1971; Goebl 1982). Looking at typology, however, it might appear at first glance that things have gone the other way round. Since the 1970s an increasing number of specific cross-linguistic properties have been studied in detail in microscopic analyses, and it has been shown that these do not at all converge with traditionally postulated holistic types, such as isolating, agglutinating, inflectional and polysynthetic. However, at about the same time the idea of holistic types was (almost) entirely aban-

doned, typologists became aware of macro-areal trends in typological feature distributions (Dryer 1989; Nichols 1992). Dryer and Nichols both pointed out the relevance of statistical approaches to typology. The discovery of macro-areal patterns of continental and hemispheric size gave rise to areal typology. All of a sudden typology had turned into a kind of global dialectology and typologists started drawing maps and compiling databases very much like dialectologists (WALS 2005).

In determining the set of varieties to be considered, dialectology, register analysis and typology have a common ideal, that of covering the full range of variability. However, in practice, this means that every approach has a very specific idea of what the most important dimensions of variation are. Register analysts strive to use corpora approximating the range of situations and communicative purposes that exist within the domain of speech and writing in a language (Biber 1985: 341). For dialectologists the areal dimension is the most important. Dialects from a whole geographical area are considered (sometimes leaving out dialects that are not areally contiguous, such as Alaska when considering the dialects of the United States, as in Grieve's contribution) with a nearly constant density of sampling points. Often, the area is divided into historically relevant subareas. Typologists tend to sample languages which are as independent of each other as possible, which means that genealogical and areal diversity is controlled for. This leads to a preference for isolates, such as Basque, in typological samples. Interestingly, the ideal of covering the full range of variability often has the consequence of a rather uniform dataset. Many register analysts of English use the British National Corpus, dialectologists often use dialect atlas data, and typologists have a battery of reference grammars as their favourite data source. Actually, if we keep the full range of linguistic diversity in mind, there are a large number of potential dimensions of variability, yet virtually all approaches in variationist linguistics restrict attention to only one or two dimensions while keeping other dimensions as constant as possible. Somewhat counter-intuitively, then, homogeneity of data on some dimensions is equally important in variationist linguistics as covering the full range of variability. Put differently, the choice of data is always driven by keeping something constant.

Not only are "ideal" datasets often astonishingly uniform, they are sometimes also highly pre-processed, which may affect the granularity of data that can be extracted. Kretzschmar (2009) notes that dialect atlases usually do not consider the full range of variation attested. Not only are dialect atlas data incomplete, they also suggest categoriality where there is none. Szmrecsanyi (2013) points out that the signal provided by conventional dialect atlases is categorical and exhibits a high level of data reduction. Wälchli (2009) makes

a similar point concerning the typological atlas (WALS 2005). The advantage of using primary data is that feature values on a lower level of data reduction can be extracted.

To the extent that feature values are determined with higher numeric accuracy (i.e., "measured" rather than determined) the question arises as to how representative the values are of the varieties they stand for. The traditional varieties "language", "dialect" and "register" are extremely broad and typologists, dialectologists, and register analysts sometimes tend to neglect that there is a great deal of language-internal, dialect-internal, and register-internal variation. This is most manifest in typology where it is quite common to compare written European standard languages with spoken language varieties elsewhere in the world. This has been criticized, for instance, by Miller and Weinert (1998) and in many papers in Kortmann (2004). In order to account for the fact that "languages" cannot be accessed directly, Wälchli and Cysouw (2012) replace it by the notion of "doculect". A doculect is any documented language variety, be it as raw data (e.g., a sound file), primary data (e.g., a transcribed text or wordlist), or secondary data (e.g., a glossed text or a grammatical description) of whatever size. In the approaches taken in this volume, "languages" and "dialects" are mostly understood as particular texts or corpora written or spoken in a particular language or dialect. A consequence of more precise measurement is that the variety the measurement stands for has to be characterized in more narrow terms; otherwise language- or dialect-internal variation may be too large in comparison to the level of measurement.

Another question is how the traditional notions of language, dialect and register relate to idiolects, i.e. to individual speakers. Typologists and dialectologists who consider data from one speaker treat languages and dialects as if they were one speaker. In some cases of moribund languages, this cannot be avoided. The documentation of the native North American language Takelma by Sapir (1909) rests on data from the single speaker alive when Edward Sapir documented the language. But even where many speakers are available typologists and dialectologists often tacitly assume that the variety-internal diversity is negligible in comparison to the cross-varietal diversity. By using the notion of doculect or by specifying the particular texts or corpora considered it can be made more explicit that a certain study focuses on cross-varietal diversity while not considering variety-internal diversity. The role of individual speakers in a dialect is considered by Corrigan, Moisl and Mearns in this volume.

Up to now we have concentrated on the notions of dialect, register, and language in this section. These three dimensions do not, however, in any way

exhaust the full range of linguistic variability, and many papers in this volume make reference to other kinds of lects. Aside from dialects we have regiolects, and Heeringa and Hinskens consider to what extent dialects may evolve into regiolects. Heeringa and Hinskens also use the sociolinguistic variables age and gender for an indirect measurement of diachronic change. Languages spoken in several countries and especially languages with written standards in more than one country are often referred to as pluricentric. National varieties in this volume are studied by Diwersy, Evert, and Neumann and by Geeraerts, Peirsman, and Speelman. Translation vs. original is another important dimension of variation considered in Diwersy, Evert, and Neumann and translations play a dominant role in the several typological contributions based on parallel texts (Dahl; von Waldenfels; Verkerk; and Wälchli). Ruette, Geeraerts, Peirsman, and Speelman emphasize the importance of taking into account several dimensions of variation at the same time, which is definitional for their sociolectometry approach. While corpora in register analysis are designed with the aim to be representative of a language in general, there is still variation across different corpora of a single language. Koptjevskaja-Tamm and Sahlgren show that lexical-typological analyses with word-space modelling can yield considerably different results when applied to different large corpora of one and the same language. On the other extreme, Mayer shows that for some universal tendencies it can be useful to use data from a large number of world languages as a single corpus. The universal characteristics of texts are emphasized most notably in Köhler's contribution dealing with linguistic laws. In linguistic laws, variations are reflected as empirical parameters. The values of such parameters usually have to be estimated on the basis of data, which is why quantitative linguistics has a strong empirical component even if its main focus is theoretical.

The variation phenomena considered in the present volume are refreshingly diverse. The volume starts with five studies whose focus is dialectal variation. First we have three dialectometric contributions. Heeringa and Hinskens explore lexical, morphological, and phonetic variation in recent dialect encodings of Dutch in the Netherlands and Flanders in 86 villages. In their study dialect change is measured in apparent time by comparing data from older male speakers with younger female speakers. Grieve considers phonetic variables in American English from the acoustic data from the Atlas of North American English from 402 informants representing 236 cities. Szmrecsanyi investigates morphosyntactic variation in 34 traditional British English dialects. Next comes a study focusing on the phonetic variation in one dialect. Corrigan, Moisl, and Mearns investigate the internal variation in Tyneside English across real time (starting from the late 1960s).

The variation among and between speakers of Newcastle and Gateshead is compared. Finally, Kretzschmar's programmatic contribution draws on the full range of lexical and phonetic variation data in the American Linguistic Atlas and the Linguistic Atlas of the Middle and South Atlantic States.

In the three very different papers by Diwersy, Evert, and Neumann, by Ruette, Geeraerts, Peirsman, and Speelman, and by Koptjevskaja-Tamm and Sahlgren variation across corpora is one of several ingredients, even though all three approaches are quite different from traditional register analysis. Diwersy, Evert, and Neumann present two case studies, one on functional variation in English and German (translations vs. originals across different registers) and one on regional variation of French across six countries. Ruette, Geeraerts, Peirsman, and Speelman's sociometric approach makes use of a large hyper-corpus of contemporary Belgian and Netherlandic Dutch to explore dimensions of lexical variation across register and national varieties of Dutch. Koptjevskaja-Tamm and Sahlgren use word-space modelling drawing on the distributional behaviour of words in three large English corpora for the sense exploration of temperature expressions.

The four contributions by Dahl, von Waldenfels, Verkerk, and Wälchli all investigate cross-linguistic variation but use very different samples of languages. All papers make use of parallel texts. Dahl explores past tense expressions, notably the perfect tense, in thirty translations of the New Testament across nine languages of Europe. One aim of the paper is to study the interplay of cross-linguistic and language-internal variation. Waldenfels explores verbal aspect and reflexive marking in ten standard varieties of Slavic with the aim of disentangling genealogical and areal factors of variation. The Slavic languages are genealogically closely related, which means this contribution in intragenealogical typology falls somewhere in between dialectology and typology. Verkerk investigates construction types used in motion events in 16 Indo-European languages; her sample is purposely restricted to languages of one language family. It is thus intragenealogical typology as well, but on a larger scale. Wälchli explores cross-linguistic variation in the encoding of lexical and grammatical domains using various world-wide samples (both balanced and convenience). The typological group of papers ends with Mayer's contribution, which is the most massively cross-linguistic paper in the volume. He considers consonant distributions in a cross-linguistic database containing word forms from over 4,300 languages in order to infer place of articulation features from sound distributions. The same method is then applied to a range of different material from various languages.

The volume concludes with the quantitative contribution by Köhler, which is concerned with the systematic relationship of variation in texts

within and across languages, as far as they can be captured in terms of linguistic laws. The paper develops a synergetic-linguistic model from which various hypotheses can be derived. One of these hypotheses concerns the relationship between word length and polysemy, which is tested on the German LIMAS corpus.

3. Corpora of text and speech

While the use of corpora is customary in register analysis and sociolinguistics, corpora are not a mainstream data source in dialectology and typology. A major aim of this volume is to illustrate the wide range of ways in which texts can be made use of for studying cross-linguistic and language-internal variation.

The most crucial contrast in the present volume is between studies that draw on naturalistic, non-parallel corpora and those that draw on parallel corpora.

Studies based on naturalistic, non-parallel corpora. Ruette, Geeraerts, Peirsman and Speelman draw on corpora sampling conversation, usenet discussion, newspaper prose, and legalese in numerous varieties of Dutch; Grieve taps into a corpus of acoustic production data from the *Atlas of North American English*, based on recordings of linguistic interviews with 439 informants; Corrigan, Moisl, and Mearns rely on a spoken dialect corpus of interviews with residents of Tyneside and surrounding areas of North East England; Diwersy, Evert, and Neumann investigate corpora sampling different registers in English and German, and a massive corpus archive of francophone newspapers; Szmrecsanyi embarks on an analysis of a spoken dialect corpus of interviews with speakers in 34 counties all over Great Britain; Koptjevskaja-Tamm and Sahlgren's study is empirically based on a large representative corpus of English (the *British National Corpus*), a newswire corpus, and a blog corpus; and Köhler draws on various corpora, mostly written.

Studies based on parallel corpora. Wälchli investigates a parallel corpus of Bible translations in 40 different doculects; Heeringa and Hinskens explore a parallel corpus that samples translations of the Chaplin movie *The Kid* in 86 dialects of Dutch, plus Standard Netherlandic Dutch and Standard Belgian Dutch; Verkerk studies a parallel corpus of translations of three novels (including, e.g., *Alice's adventures in Wonderland*) in 16 languages; Dahl analyses a parallel corpus of Bible translation in 9 languages; and von Waldenfels examines a parallel corpus consisting of a novel and its translation into 11 Slavic languages as well as into Modern Greek and German.

The two studies in this volume that do not neatly fall in either camp are Mayer's, which explores a cross-linguistic database containing word forms from over 4,500 languages, and Kretzschmar's, who presents case studies based on quite diverse material. It is obvious, though, that the parallel-corpus-based studies in this volume tend to be cross-linguistic in nature (exception: Heeringa and Hinskens), while the naturalistic-corpus-based studies tend to be concerned with language-internal variation (exceptions: Koptjevskaja-Tamm and Sahlgren, and Köhler). Evidently, the choice of parallel corpora is driven by the need to keep propositional content constant – a more pressing need in cross-linguistic research, where it is harder than in research on language-internal variation to keep other things such as, e.g., surface form constant.

Another important distinction, of course, is the one between studies relying, wholly or partially, on corpora sampling *spoken* material (Ruette, Geeraerts, Peirsman, and Speelman; Grieve, Corrigan, Moisl, and Mearns; Heeringa and Hinskens; Kretzschmar; Szmrecsanyi; and, to the extent that the *British National Corpus* samples spoken material, Koptjevskaja-Tamm and Sahlgren), and studies analyzing *written* material (Wälchli; Diwersy, Evert, and Neumann; Verkerk; Dahl; Köhler; von Waldenfels). Again, it is evident that cross-linguistic studies tend to analyse written data (exception: Koptjevskaja-Tamm and Sahlgren), while studies concerned with language-internal variation typically rely on spoken material (exception: Diwersy, Evert, and Neumann). Notice, in this connection, that parallel corpora are not necessarily written – Heeringa and Hinskens' parallel corpus is, in fact, spoken.

4. Features

In linguistics, the answer to the question "What is a feature?" can vary a great deal across sub-disciplines. Features are complex units and there is much inconsistency in the literature about what aspects of features are subsumed under the notion of "features". We find it convenient to introduce the three terms *box* (where a feature can occur), *status* (by what value the feature is instantiated), and *resulting value* (how the feature value is computed) to highlight the complex nature of features. The terms are inspired by graphical user interface terminology which makes them maximally neutral in linguistics, not giving precedence to any existing linguistic framework. To give a very simple example: if we determine the type-token ratio in a text, each word form token is a box and the status of the box is the word form type. In order to obtain the resulting value, the type-token ratio, we have to apply a certain operation to the set of statuses across the set of boxes, in this case simply

dividing the number of boxes by the number of statuses. Box, status and resulting value are closely akin to the traditional statistical notions "variable" (the thing measured) and "value" (the measurement). However, while box only corresponds to variable, both status and resulting values are values, but they are rather different kinds of values, which makes it important to have two different terms for them.

In this chapter we shall make use of the term "feature" in two different ways: (i) when we discuss how the term is used in various approaches (in this case, we put the term in quotation marks); or (ii) when we refer to the box/status/resulting-value complex as such (for instance, when we discuss whether a certain study considers a phenomenon to consist of one or several features), which is in our view the most appropriate use of the term "feature". In the foregoing discussion we have sometimes used the compound *feature value* when emphasizing the resulting-value component of features.

Formal approaches to features, such as Carpenter (1992), are concerned with issues such as defining complex hierarchies and inheritance patterns of boxes and statuses ("feature structure"). In the corpus-based studies in this volume, however, the complexity of features arises from their multiple instantiations in texts. Boxes and statuses are sets of exemplars rather than just single abstract units in the language system. Corpus-based studies will thus usually want to relate the sets of boxes and statuses in some quantitative manner, which is why resulting values are an indispensable third aspect of features in most studies in this volume. Resulting values give features a more procedural character than they have in formal approaches. Let us now first consider how features are treated in traditional approaches to sociolinguistics, register analysis, and typology.

Labovian sociolinguistics and many dialectological studies work with the notions of *variable* (box) and *variant* (status). We may, for instance, be interested in whether and how /r/ is pronounced in a certain variety of English in certain positions in a certain set of words. The variable (r) can, for instance, have two variants, [r] and Ø, in postvocalic contexts. Note that sociolinguistic variables are often phonetic, but the same approach can be applied to grammatical or lexical variables. Building on Labov (1972: 271), a sociolinguistic variable is typically defined such that the variants of the variable must be different ways of saying the same thing. Counting the frequency of a variant vis-à-vis competing variants is how sociolinguists typically determine resulting values.

In register analysis à la Douglas Biber (1985 and elsewhere), the units called "features" are morphosyntactic and lexical properties that can easily be identified in texts, using automatic and semi-automatic retrieval techniques.

Table 2. Typical treatment of features in various approaches.

	Box	Status	Resulting value
Sociolinguistics	Variable	Variant	Frequency (proportion) of variant per variable
Register analysis	Word form	"Feature"	Frequency of "feature" e.g. per million words
Typology	Domain	Type	Dominant type in a domain

Examples in English include past tense forms, first and second person pronouns, and time adverbs. In order to determine what kinds of units these properties are it is useful to look at the resulting values first. Typical resulting values are the frequencies of properties per number of word tokens (or per larger text units, such as the number of sentences). Hence, the boxes are the word forms (or sentences): every word form (or every sentence) is viewed as a possible environment for the occurrence of a property, and the properties (the "features") are statuses.

Unlike sociolinguistics and register analysis, typology is cross-linguistically comparative. Therefore, the units to be compared cannot be language-particular words, grammatical constructions, or sounds. Instead, it is common to choose semantic units that are encoded in all languages as comparative concepts. These semantic units are also called *functional domains*. For instance, we can consider the way in which a language expresses past events and whether a past tense is used to express them (irrespective of the way in which the past is formally marked), and whether it makes any distinctions regarding remoteness (such as hodiernal vs. remote past). This results in a typology consisting of (1) languages with no past tense, (2) languages with a single past tense, and (3) languages with remoteness distinction in their past tenses. These types are statuses, and the domain (here: the expression of past events) is the box. It is quite common in typology to equate types (statuses) with "features", as in traditional implicative universals: A language that has "feature" *p* will also have "feature" *q*.

Table 2 summarizes the typical treatment of features in the three traditional approaches found in this volume.

Sociolinguistics and register analysis, on the one hand, and typology, on the other hand, differ in whether features are selected just because they are indicative of cross-varietal variation or whether they are interesting objects of study in themselves. Typologists are usually interested in exploring the nature of features, while sociolinguistics and register analysts mostly choose features because they may serve to characterize varieties and/or speech

communities. Put succinctly, sociolinguistics and register analysts are mainly interested in the distribution of features; typologists are mainly interested in the features themselves. Dialectologists may pursue both goals. In any event, the dialectological, sociolinguistic, and register analytic papers in this volume emphasize the importance of considering many features in order to characterize a variety. The typological papers in this volume, however, do not focus mainly on the joint characterization of whole languages, notably because correlations across different domains are much weaker in distantly related languages than in dialects and registers, which makes it less promising to apply dimension reduction across domains in typology. However, using aggregation of multiple features is particularly useful for exploring intra-domain variability in typology.

There are many ways, then, in which features and their interpretation can vary within and across approaches. Relevant dimensions include level of measurement, ontological mode and degree of supervision in feature selection.

Since both statuses and resulting values are values from a statistical point of view, they can differ in their *level of measurement,* or scale of measure: they can be binary, nominal/categorical, ordinal, interval, or ratio. The level of measurement is particularly important for resulting values. The frequencies customary in sociolinguistics and register analysis are ratio scales. Traditional typological "features", however, are often nominal or ordinal. For instance, the past tense typology mentioned above can be viewed both as nominal (three different types) or ordinal (a scale ranging from no to one and more than one past tenses). However, the fact of choosing a nominal or an ordinal scale in typology has more to do with tradition than necessity. For instance, word order is usually determined categorically as the dominant word order in a language. Yet word order can be equally well measured in texts as a variant per variable. Many resulting values in typology are distributed bimodally when measured as ratio scales. This invites data reduction to nominal scale which, however, entails a considerable loss of data. Statuses are most often nominal/categorical except for the binary "features" in phonology (and other formal approaches), such as sonorant or continuant, which can only assume the statuses "plus" and "minus".

When interpreting features, one must take into account whether these are viewed as essential units or convenience units. This can be called the *ontological mode* of features. While essential features are often "determined" and convenience units generally "measured", it is important to bear in mind that the ontological mode is a matter of interpretation, not of measurement. The question is whether we believe that features are underlying properties of the

language variety they characterize, or whether we believe that they are just observable surface phenomena. Put differently, is a feature status the underlying structural parameter in a language system, or is it a measurement in a speech or text sample? It is often wrongly believed that essential units can be obtained from convenience units by data reduction from interval or ratio scale to categorical scale. Phonological "features" are a typical example of a kind of feature that is widely believed to be an underlying essential unit. Most papers in this volume have in common the fact that they view features as convenience units rather than essential units. This follows quite naturally from the focus on text in this volume.[2] Quantitative linguists have advocated non-essentialist approaches in typology for a long time (Altmann and Lehfeld 1973).

Another question is whether features are preselected by the researcher or whether they emerge in the course of investigation without supervision. Without going into much detail, it is important to note that the notion of supervised vs. unsupervised is not trivial at all. There are almost no truly unsupervised approaches since we cannot really obtain values for features without programming for them, or counting by hand. In both cases the features are indirectly or directly determined by the researcher. For our purposes it is therefore more important to consider whether features are given or emerge in the course of the investigation (whether or not this is called "unsupervised"). The advantage of emergent feature selection is that the results of an aggregation will be less predetermined by the researchers' expectations. However, one disadvantage is that many irrelevant features are considered (which can be corrected by manual processing steps). In this volume, emergent feature selection is addressed in Diwersy, Evert and Neumann and in Ruette, Geeraerts, Peirsman and Speelman.

Drawing on the box/status/resulting value parlance introduced above, the papers in this volume can be characterized as follows:

Heeringa and Hinskens' study considers lexical, morphological and phonetic distances in a parallel text corpus consisting of maximally 125 words per text. In terms of features every one of the maximally 125 words of a text is a box. For lexical comparison, the lexeme is the variant; thus s*nel*, *rap* and *gauw* are three variants (statuses) for 'quickly'. For morphological compari-

[2] The recent monograph on features by Corbett (2012) treats features exclusively as essential entities in a traditional structuralist and formalist perspective and focuses on morphosyntactic features, such as number and gender, which are more often referred to as morphosyntactic categories elsewhere. Corbett's features are boxes, the "values" are statuses, and resulting values do not figure since his approach to features is not procedural.

son, each word with a different derivation and/or inflection is considered a different variant. For the sound components, Levenshtein distance and alignment lengths are used to measure the phonetic distance between varieties in a pairwise fashion, taking words as the basic analytical units. Thus, every form with some transcribed difference is a variant of its own. However, no resulting values are determined prior to the pairwise-aggregational comparison across varieties.

Grieve remains close to Labovian sociolinguistics in the way features are defined. The boxes are formant 1 and formant 2 of 38 vowels from the acoustic data of the *Atlas of North American English* (Labov, Ash and Boberg 2006). Statuses are the formant values and the resulting values are the average formant values per dialect.

In his study of British English dialects, Szmrecsanyi combines elements of both the sociolinguistic and the register analytic approaches. He defines 57 morphosyntactic "features", measures their text frequencies in the corpus (*log*-transformed frequencies per 10,000 words), and subsequently explores the "feature" portfolio's joint frequency variability as a function of language-external parameters. As in register analysis, Szmrecsanyi's "features" are statuses and the resulting values used for aggregation are the frequencies of variants. However, unlike in register analysis, comparability across varieties is assured by choosing texts of the same register, oral history material from the *Freiburg Corpus of English Dialects*.

Corrigan, Moisl and Mearns compare what they call "feature-based" and "aggregate" analyses of the *Diachronic Electronic Corpus of Tyneside English* (DECTE). By "feature"-based analysis they refer to the traditional sociolinguistic approach where the GOAT vowel, relative clause marking and intensifiers are used as variables. In the "aggregate" analysis part, Corrigan, Moisl and Mearns compute a data matrix with 64 speakers in the corpus and 156 phonetic elements as the two dimensions. For each phonetic element they consider how often it occurs in the production of individual speakers in the corpus. These phonetic elements are "features" in the sense of statuses. The boxes are the number of phonetic segments. The resulting values are obtained by length normalization.

According to Kretzschmar, the formal notion that a dialect or language "has" some particular "feature" (in the sense of variant of a variable, i.e. a status) becomes untenable once we pay attention to the massive variation in all "features" (in the sense of a variable, i.e. a box) in all language varieties. When the variant types of any linguistic feature box are graphed according to their token frequency, the chart typically exhibits a nonlinear asymptotic hyperbolic curve characterized by a small number of highly frequent responses

and a much larger number of less-frequently-occurring responses. This speaks against simply determining a dominant variant of a variable as a resulting value to characterize a feature (as is common in dialect atlases). Another way to put this is to say that language structure is emergent.

Diwersy, Evert, and Neumann present two case studies, one on functional variation in English and German (translations vs. originals across different registers), and one on regional variation in French in six countries. The first case study follows the standard setup of register analysis with 29 lexicogrammatical "features" (statuses) where boxes mostly are word tokens and sentences. The "features" are chosen such that they can be identified both in English and German. For the resulting values, ratios are transformed to standardized z-scores due to strikingly different value ranges of the ratios. The second case study uses 8,248 pairs of functionally disambiguated nouns with their passive or active valency collocations (lemma and syntactic-functional label). The resulting values are again based on frequency. A major difference between the two case studies is that the "features" (statuses) are given in the first study while they are extracted in an emergent fashion in the second study.

In Ruette, Geeraerts, Peirsman and Speelman, the variables are lexical sets of synonyms which are first selected automatically without supervision by a Clustering by Committee algorithm in a large Belgian and Netherlandic Dutch corpus with texts from different registers. The variables (boxes) are the 200 sets of synonymous nouns retained after manual clean-up and the variants (statuses) are the individual nouns (the synonyms in the synonym sets). The resulting value, or rather the set or distribution of resulting values, is the profile, viz. the set of frequencies of each variant in a variable which then serves as a basis for aggregation. Ruette, Geeraerts, Peirsman and Speelman argue that lexical variables and profiles are the same thing. However, in terms of our distinction of box and resulting value there is a clear difference: the variable is the box and the profile is the complex of resulting values. Boxes (variables) and resulting values go quite naturally together in this approach, because variables are not given from the outset but rather are determined in the course of the investigation. Despite their register analytic approach, Ruette, Geeraerts, Peirsman and Speelman's conception of features is rather close to Labovian sociolinguistics, and the paper explicitly refers to Labov's notion of the linguistic variable as their starting point.

In Koptjevskaja-Tamm and Sahlgren's study of English temperature expressions, the boxes are the temperature words to be investigated and the statuses (or "features") are the contexts in which these words occur. The context "features" (i.e. observations of co-occurrence events) are then ag-

gregated to word space models. The aim is to explore the semantics of the compared words (assuming that similarity of distribution and similarity of meaning correlate). The use of different corpora has a heuristic function, that is, to explore how balanced a corpus has to be in order to be useful for the method and to investigate the context sensitivity of words. The aim is to aggregate words, not corpora. The results are compared to findings in lexical typology.

Dahl, Verkerk, von Waldenfels, and Wälchli are all typological investigations using distributions across parallel texts for feature aggregation. The boxes are the aligned contexts across the parallel texts, which are similar, but not identical, in meaning and which, in their entirety, constitute a functional domain. Dahl investigates past tense expressions, notably the perfect, Verkerk motion events, von Waldenfels aspect and reflexives and Wälchli negation and nominal number as well as some lexical domains. In contrast to the register analytic, dialectological and sociolinguistic papers in this volume, the much narrower focus of these papers (focusing on one or a few functional domains) is well in line with the observation that typology is as much interested in the particular nature of features as in the configuration of language varieties emerging from the aggregation of features. In all four papers, the boxes – the aligned contexts in the parallel corpora – are filled by language-particular form classes with similar but not identical functions. As in traditional typology, the language-particular form classes can be assigned to types such as the perfect gram type in Dahl. Where the languages to be compared are closely related, the language-particular form classes are sometimes genealogically related, such as the perfective and imperfective aspect and the reflexive in Slavic in von Waldenfels' study. Verkerk investigates motion events in 16 Indo-European languages and classifies examples according to nine different construction types. In all three papers the resulting values are difficult to distinguish from aggregation discussed in Section 4. Even though the contexts in the parallel texts express the same domain, every context (box) functions as a feature of its own which carries a certain categorical nominal value (status). These values can be compared for the purpose of aggregation across varieties to compute (dis)similarity matrices of cross-linguistically similar grams based on their similarity of distribution. Wälchli partly makes use of a similar methodology, but also measures the distributional similarity of language-particular form classes without assigning them to cross-linguistic types by using pair-wise comparison of distribution similarities of language-particular form classes.

It should be noted that the features in the four typological papers using parallel texts are quite different from "features" in traditional typology. The

feature is not a general domain, such as past tense reference; rather, each individual context in the parallel text is a feature of its own. These features are all of the same kind and they can be arranged in a map, as in Dahl's paper, where the box has a constant position in the map and the status or resulting value is marked by colour. While it is not possible to aggregate over a single domain in traditional typological investigations, aggregation over the many contexts that make up a domain is easily possible in typological approaches based on parallel texts and serves to explore a single typological domain in a bottom-up fashion, as emphasized by von Waldenfels.

Mayer deals with the inference of distinctive phonological "features" from the distribution of sounds. Rather than identifying phonological "features" in terms of articulatory and acoustic properties, they emerge by way of hierarchic clustering of sounds based on their distribution in word lists and in texts. The paper focuses on place of articulation, which is argued to be conditioned by a universal tendency of Similar Place Avoidance.

Köhler addresses features from the perspective of Synergetic Linguistics. He starts by asking why semiotic systems change. Every time new variants appear, the survival probability of the resulting configurations of "features" (i.e. the variants or statuses competing for certain communicative functions: the variables or boxes) depends, among other things, on how well they succeed in achieving the intended communicative purpose. All factors involved mediate, via the so-called *order parameters*, between the needs of language users and the mechanisms of production and perception. An example of such a need is the requirement of minimisation of production effort (Zipf's [1949] "principle of least effort"). Synergetic Linguistics suggests that features should not be considered in isolation. Köhler further emphasizes that features under study must be transformed into quantitative variables (i.e. metrified). In other words, resulting values of features must be determined.

At the end of this section, let us summarize some general trends and point out the outliers. While most papers in this volume go from text via features to aggregation across varieties – and this holds equally for dialectology, sociolinguistics, register analysis, and typology – Mayer uses aggregation to determine features in an unsupervised fashion. The features to be found by partitioning are phonological features which are predicted to be underlying essential features. Most approaches represented in this volume – and this holds again equally for dialectology, sociolinguistics, register analysis, and typology – determine resulting values which are commensurable across dialects, registers and languages. However, Wälchli and Heeringa and Hinskens use pair-wise comparisons to make feature statuses commensurable. The use of Levenshtein distance to calculate pronunciation differences between

words (which is fairly popular in dialectometry) operates via pair-wise comparison. Wälchli uses pair-wise similarity in distribution between forms without classifying forms into cross-linguistic types in exploring the domain of negation.

A major difficulty in focusing exclusively on features (as this section has sought to do) is that in many approaches features and aggregation are so intimately intertwined that it is hardly possible to keep them apart. Let us now delve into the issue of aggregation.

5. Aggregation and quantitative analysis techniques

Aggregation – as we understand the term here – is the joint analysis of multiple features. Aggregation almost necessarily involves usage of some quantitative analysis technique to (i) synthesize data on individual features, and (ii) to analyse the aggregate dataset. That said, aggregation can have various purposes. It can serve to characterize a language or language variety as a whole, and to obtain a more robust and more reliable linguistic signal than would be possible in a single-feature based study. These objectives are particularly important in dialectology, register analysis and – albeit to a lesser degree – sociolinguistics. In the present volume, the contributions by Heeringa and Hinskens; Grieve; Szmrecsanyi; Corrigan, Moisl and Mearns; Diwersy, Evert and Neumann; and Ruette, Geeraerts, Peirsman, and Speelman fall into this camp. However, multiple-feature approaches can also be used to create a fine-grained analysis of a larger domain, which is the approach taken in most typological contributions in this volume (Koptjevskaja-Tamm and Sahlgren; Dahl; von Waldenfels; Verkerk; Wälchli; and Mayer). Kretzschmar's and Köhler's contributions are mainly concerned with theoretical issues.

On the technical plane, the papers in the present volume that have an empirical focus (that is, all except Kretzschmar's and Köhler's) use different methods to aggregate features, and to analyse and/or visually depict the outcome of the aggregation endeavour. There are largely two families of aggregation techniques used in the volume. These can be characterized broadly as one-step and two-step procedures.

One-step procedures include Factor Analysis (FA) and Principal Component Analysis (PCA). Both produce ranked lists of latent dimensions, or put differently, they reduce the *n* dimensions of the *n* features considered to a smaller number of dimensions profiting from latent inter-dependencies between features. PCA is used especially with large numbers of features in relation to the number of varieties studied, a scenario in which FA fails to con-

verge (see Diwersy, Evert and Neumann for a comparison of FA and PCA). Grieve uses FA in multivariate spatial analysis. An advantage of one-step procedures is that the contributions of particular features to resulting dimensions can be traced more easily. Both FA and PCA are basically unsupervised. Diwersy, Evert and Neumann use Linear Discriminant Analysis (LDA) for semi-supervised aggregation which can be combined with selected unsupervised principal components.

In two-step procedures the first step is the calculation of a distance or (dis)similarity matrix, which is a two-dimensional array of the size $N \times N$ containing the distance or dissimilarity between any pairs of N units, much like e.g. distance tables to be found in road atlases (see von Waldenfels' contribution for a concrete illustration). Distance matrices are usually calculated from an $N \times p$ feature matrix by using a particular distance measure. Distance measures used in this volume are Hamming (Dahl, von Waldenfels, Verkerk), Euclidean (Szmrecsanyi), City-Block or Manhattan (Ruette, Geeraerts, Peirsman and Speelman), Jaccard/Dice (Wälchli), T-score (Wälchli), the Phi coefficient (Mayer), and Levenshtein (Heeringa and Hinskens). For the details we refer to the individual papers and the references given there, but generally it can be said that Hamming is the most general measure for nominal/categorical data and Euclidean is the simplest measure for interval or ratio scale data. Jaccard/Dice and T-score are simple collocation measures (Manning and Schütze 1999: 151–189). For assessing the degree of positive or negative collocation of features across varieties, Spearman's squared Rho can be used (Wälchli).

Another possibility is to construct a vector space and to measure the similarity of units with a vector similarity metric. Word Space models (Koptjevskaja-Tamm and Sahlgren) calculate meaning similarity between words on the basis of the contexts in which they occur and represent it as proximity in high-dimensional vector spaces. The simplest vector similarity metric, used by Koptjevskaja-Tamm and Sahlgren, is the cosine metric. Vector space models are also used by Ruette, Geeraerts, Peirsman, and Speelman, but not for the purpose of aggregation.

Since high dimensionalities are hard to grasp when one merely eyeballs the data, what is needed are sparkling visualization techniques that reduce the high dimensionality of distance matrices. Standard methods for this purpose are hierarchical agglomerative clustering visually represented in dendrograms (Mayer; Corrigan, Moisl and Mearns) and Multi-dimensional scaling (MDS). Like Factor Analysis, MDS serves to reduce the dimensionality by rearranging the data points such that as much information as possible can be captured by few dimensions. There are several varieties of MDS, the sim-

plest being classical multi-dimensional scaling. Ruette, Geeraerts, Peirsman, and Speelman use Kruskal's non-metric MDS. Network diagrams, which are essentially multiple dendrograms, are another important visualization device. Neighbor-nets (Szmrecsanyi; Dahl; Verkerk; von Waldenfels; Wälchli) are useful for visualizing messy datasets supporting more than one tree analysis. The shortest path between two nodes along the net is proportional to their distance in the matrix. Neighbor-nets can be combined with bootstrapping (von Waldenfels). An alternative to hierarchical clustering is partitioning. Koptjevskaja-Tamm and Sahlgren use syntagmatically labelled partitioning. Various geographical projections of distance matrices can be used. For instance, MDS dimensions can be mapped on the red-green-blue colour scheme. Szmrecsanyi demonstrates the pros and cons of beam maps, similarity maps, cluster maps and continuum maps.

In addition to the most fundamental processing steps, there are further procedures (such as weighting and smoothing) for adapting features before aggregate analysis and for verifying the consistency of matrices. Ruette, Geeraerts, Peirsman and Speelman use weights to ensure that concepts which have a relatively higher frequency also have a greater impact on the distance measurement. Heeringa and Hinskens use graded weights to measure distance in sound components. The simplest approach is to weight all features equally. In Grieve's multivariate spatial analysis the individual features are initially smoothed using a local spatial autocorrelation analysis so as to identify underlying patterns of spatial clustering in the values of each variable. Cronbach's alpha can be used to measure the internal consistency of the features.

Let us now outline the aggregation techniques used in the individual papers:

Heeringa and Hinskens use corpus-derived features to generate a number of distance matrices. These distances are subsequently projected to geography in various maps, and/or analysed using standard inferential statistics. In addition, Heeringa and Hinskens marshal an exploratory statistical analysis technique: Hierarchical Agglomerative Cluster Analysis, or Cluster Analysis for short (here and in the following the reader is referred to the actual contributions for definitions and explanations of the techniques used).

The contribution by Grieve uses an impressively long though well-motivated line-up of statistical analysis techniques: the corpus-derived features (vowel measurements in this case) are the input for a Local Spatial Autocorrelation Analysis, that output is passed as input to Factor Analysis, which is an exploratory statistical analysis technique that serves as the first aggregation step in the study. Next, Grieve uses the output of Factor Analysis to calculate a distance matrix (the second aggregation step), which is then

passed on to a Cluster Analysis. Results are visually depicted primarily via various projections to geography.

Szmrecsanyi uses the Euclidean distance measure to generate an aggregate distance matrix from the usage frequencies of the features that are considered in the study. On the basis of this distance matrix, Szmrecsanyi generates various projections to geography, as well as a Neighbour-Net diagram, a visualization technique originally developed to address evolutionary biology issues and now also often used to gauge relationships between languages or dialects. Szmrecsanyi moreover explores aggregate linguistic distances using Multidimensional Scaling (yet another exploratory analysis technique), and draws on regression techniques to quantify the relationship between geographic and aggregate linguistic distances.

Corrigan, Moisl and Mearns also use a distance matrix-based aggregation approach but calculate linguistic distances between dialect speakers – not dialects – using frequency data for a range of phonetic segments that these speakers use; we note that the extraction of this data from the corpus material requires substantial amounts of data transformation prior to aggregation. The distance matrix created by Corrigan, Moisl and Mearns is analysed using Cluster Analysis, and the results are visually depicted in dendrograms, which are in turn correlated with social data.

Diwersy, Evert and Neumann present a sophisticated 7-step methodology to apply semi-supervised multivariate techniques to datasets comprising many corpus-derived variation phenomena. Their methodology involves Principal Component Analysis (an exploratory statistical analysis technique related to Factor Analysis), Linear Discriminant Analysis (which is less exploratory than Principal Component Analysis), and supervised machine learning with Support Vector Machines. Visualization occurs at various stages of the aggregation endeavour via scatter and line plots. Crucially, Diwersy, Evert and Neumann's aggregation technique is not distance matrix-based.

Ruette, Geeraerts, Peirsman, and Speelman are concerned with creating lexical distance matrices. What is particularly innovative here is that the feature portfolio is created bottom-up in a semi-automatic fashion via Semantic Vector Space Models (which were originally developed in information retrieval and related fields). Using three distance metrics with various weighting characteristics, they calculate three distance matrices, each detailing lexical distances between lects sampled in the corpus database. To aid comparison of the matrices, the authors utilize Multidimensional Scaling to create two-dimensional plots.

Koptjevskaja-Tamm and Sahlgren use word space models for aggregating co-occurrence events. The aim is to provide an aggregated view of the dis-

tributional behaviour of temperature words in large corpora. Paradigmatic neighbours are visualized as word clouds and with a MDS representation.

Dahl first uses a two-dimensional representation of a single scale between the perfect prototype and the "anti-prototype" in past time reference, which is suitable to visualize the concrete diversity across the languages considered. He then uses a Neighbour-Net graph based on a distance matrix calculated with Hamming distance to represent the dissimilarities of all texts considered.

A very similar procedure is applied by von Waldenfels, but on the basis of two categories: aspect and reflexives. Like Dahl, von Waldenfels calculates distance matrices with Hamming distance. Bootstrapping is employed to validate the visualization in the graph. Near neighbour graphs are used to visualize the results.

Verkerk uses distance matrices calculated with Hamming distance and Neighbour-Nets in a similar vein as Dahl to identify manner- and path-salience in motion events across Indo-European languages.

Wälchli extracts word forms and morphemes in lexical and grammatical domains and uses them to calculate various indices of morphological typology visualized in world maps. Their cross-linguistic correlation is measured with the square of Spearman's Rho. The distributional diversity of the forms extracted in one domain is analysed with distance matrices computed with a simple collocation measure (Dice) and visualized as neighbour nets. The similarity of doculects in a particular grammatical domain is then measured by calculating an index of the previously measured distributional similarity of each pair of forms reflecting the category in pairs of languages of that domain.

Mayer infers place of articulation features of consonants by first calculating distance matrices of consonant pairs in CVC sequences with the Phi coefficient and then applying Cluster Analysis visualized in dendrograms.

Köhler views "linguistic laws" as means for extreme data compression, and offers that there can be no scientific explanation or prediction without laws. He distinguishes distributional, functional, and developmental laws. He goes on to sketch methods to propose universal hypotheses, which may become laws and thus a part of a theory, i.e. a system of laws. Two approaches to systems of laws in quantitative linguistics are discussed: *Unified Theory* and *Synergetic Linguistics*. Unified Theory is a mathematical approach in the form of a differential equation. It is a kind of aggregation of laws; a large number of well-known laws and hypotheses from quantitative linguistics can be derived from it. Synergetic linguistics investigates self-organising systems far from equilibrium. Synergetic linguistics can be used to model subsystems of language. The paper shows how synergetic linguistics can be used to test the hypothesis that word polysemy is a function of word length measured in morphs.

6. The structure of this volume

In the preceding sections we have presented the contributions according to the four dominant themes of this volume: linguistic varieties, corpora of text and speech, features, and aggregation with quantitative analysis techniques. However, this volume does not simply consist of variations on four themes; each paper has its own intrinsic research agenda, which goes to show that variation-oriented aggregation of features in text and speech is highly flexible – it has an enormous potential of application to different research questions in text and speech, in dialectology, register analysis, typology, and quantitative linguistics. We hasten to add that this potential is far from being exhausted by the fourteen papers in this volume.

So while the volume clearly unites approaches from different branches of linguistics, it would not be easy, nor advisable, to assign the papers to one of five boxes – dialectology, sociolinguistics, register analysis, typology, and quantitative linguistics – which is why this volume does not group the papers into sections. That said, there is a certain progression in this volume, from works in dialectology, via register analysis, to typology and synergetic linguistics.

To conclude, we would like to emphasise what the papers in this volume have in common: They are all usage-based studies in variationist linguistics, and all draw on features and aggregation methodologies – even though the diversity of the features (and what they stand for), as well as the multiplicity of quantitative analysis techniques used in these papers testify to the wide range of approaches empirical linguistics has to offer.

References

Adger, David and Graeme Trousdale 2007 Variation in English Syntax: Theoretical Implications. *English Language and Linguistics* 11: 261–278.

Altmann, Gabriel and Werner Lehfeldt 1973 *Allgemeine Sprachtypologie. Prinzipien und Meßverfahren*. München: Fink.

Atkinson, J. Maxwell and John Heritage (eds.) 1984 *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.

Biber, Douglas 1985 Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics* 23: 337–360.

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

Bybee, Joan L. 2010 Language, usage and cognition. Cambridge, New York: Cambridge University Press.

Carpenter, Robert L. 2005 *The Logic of Typed Feature Structures: With Applications to Unification Grammars, Logic Programs and Constraint Resolution*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2012 *Features*. Cambridge: Cambridge University Press.
Cysouw, Michael to appear In: Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi (eds.), *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*. Berlin, Boston: Walter de Gruyter.
Dryer, Matthew S. 1989 Large linguistic areas and language sampling. *Studies in Language* 13 (2): 257–292.
Goebl, Hans 1982 *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.
Kortmann, Bernd (ed.) 2004 *Dialectology Meets Typology: Dialect Grammar from a Cross-Linguistic Perspective*. Berlin, New York: Mouton de Gruyter.
Kretzschmar, William A. Jr. 2009 *The Linguistics of Speech*. Cambridge: Cambridge University Press.
Labov, William 1972 Some principles of linguistic methodology. *Language in Society* 1(1): 97–120.
Labov, William, Sharon Ash, and Charles Boberg. 2006 *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.
Manning, Christopher D. and Hinrich Schütze 1999 *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.
Miller, Jim and Regina Weinert 1998 *Spontaneous Spoken Language*. Oxford: Clarendon.
Nerbonne, John 2008 Variation in the Aggregate: An Alternative Perspective for Variationist Linguistics. In: Kees Dekker, Alasdair MacDonald and Hermann Niebaum (eds.), *Northern Voices: Essays on Old Germanic and Related Topics Offered to Professor Tette Hofstra*, 365–382. Leuven: Peeters.
Nichols, Johanna 1992 *Linguistic Diversity across Space and Time*. Chicago: Chicago University Press.
Sapir, Edward 1909 *Takelma Texts*. Anthropological Publications of the University Museum; Philadelphia: University Museum.
Séguy, Jean 1971 La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335–357.
Szmrecsanyi, Benedikt 2013 *Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry*. Cambridge: Cambridge University Press.
Tomasello, Michael 2003 *Constructing a Language. A Usage-Based Theory of Language Acquisition*. Cambridge, Mass.: Harvard University Press.
Wälchli, Bernhard 2009 Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13 (1): 77–94.
Wälchli, Bernhard and Michael Cysouw 2012 Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50 (3): 671–710. (Theme issue edited by Maria Koptjevskaja-Tamm and Martine Vanhove, New Directions in Lexical Typology).
WALS 2005 Martin Haspelmath, Matthew Dryer, David Gil and Bernard Comrie (eds.), *The World Atlas of Language Structures*. (Book with interactive CD-ROM) Oxford: Oxford University Press.
Zipf, George Kingsley 1949 *Human Behavior and the Principle of Least Effort*. (Reprint 1972.) Cambridge: Addison-Wesley. New York: Hafner.