D.KH. KHOLMURODOVA, N.B. NAZAROVA,
F.B. KHANDAMOVA, SH.V. KARIMOVA

CORPUS LINGUISTICS

MINISTRY OF HIGHER EDUCATION, SCIENCE AND INNOVATIONS OF THE REPUBLIC OF UZBEKISTAN

SAMARKAND STATE INSTITUTE OF FOREIGN LANGUAGES

D.KH. KHOLMURODOVA, N.B. NAZAROVA,
F.B. KHANDAMOVA, SH.V. KARIMOVA

CORPUS LINGUISTICS

MANUAL

SAMARKAND – 2024

UDC: 755.36.01

D.Kh. Kholmurodova, N.B. Nazarova, F.B.Khandamova, Sh.V. Karimova - "Corpus linguistics". – Samarkand: SamSIFL, 2024. 100 p.

The manual introduces the concepts of corpus linguistics, makes it possible to master the basics of corpus technologies, acquire skills in working with corpora, determine the place of the discipline and the corpus itself in a number of information technologies.

Designed for students, undergraduates and graduate students of philological specialties.

Reviewers: **PhD. G.H.Obrueva**

PhD. N.J.Suleymanova

The manual was recommended for the students of – 70111801 – Foreign language and literature (English language) and 70230101 – Linguistics (English language)

Ushbu uslubiy qo'llanma Samarqand davlat chet tillar instituti O'quv-uslubiy kengashining 2023-yil 13-iyuldagi 10-11-sonli qo'shma yig'ilish bayonnomasi bilan nashrga tavsiya etilgan.

© Samarkand state institute of foreign languages, 2024.

CONTENT

Foreword..... 4

1. Basic concepts of Corpus Linguistics..... 6

2. Types of Corpus..... 10

3. Technology of Corpus Linguistics..... 16

4. Areas of application Corpus Linguistics..... 25

5. Theoretical foundations of Corpus Linguistics..... 34

6. Practical basics of Corpus Linguistics..... 42

7. Creation and use of Corpus Linguistics..... 48

8. Classification of Corpus..... 52

9. The notion of discourse..... 55

10. International organizations related to Corpus Linguistics..... 60

11. Principle of text competing technologies: rule based, static constraint based restriction based hybrid..... 64

12. The notion of pragmatic intention and its types..... 71

13. Tokenization, lemmatization and parsing..... 78

14. Brown Corpus USA and LOB Corpus in Europe..... 82

15. Traditional and modern methods of linguistics research..... 86

Glossary..... 93

References..... 98

SmDCHTI ARM

FOREWORD

The manual brought to your attention is a kind of generalization of numerous disparate materials published over the past two decades in Uzbekistan and abroad, which formed the basis of lecture courses in the discipline "Corpus Linguistics".

The purpose of the manual is to acquaint students with the concepts of corpus linguistics, help them master the basics of corpus technologies, acquire skills in working with corpora, determine the place of the discipline and the corpora proper in a number of information and linguistic technologies.

Tasks of the manual:

- to familiarize students with a new paradigm in linguistic research;
- to familiarize students with the history of corpus research;
- to familiarize students with language and software tools of corpus linguistics;
- to form students' skills in working with software tools and information resources of corpus linguistics;
- to familiarize students with specific linguistic research based on corpus data.

Corpus linguistics is a new direction in linguistic science, which makes it possible to study units of any language level in their actual use, i.e., taking into account the situation in which this or that statement was made.

Large national corpora and corpora created for special purposes allow researchers to automatically search for and organize empirical material and quickly process large amounts of linguistic data.

This is one of the rapidly developing areas, and if we consider that corpus linguistics is, first of all, a methodology for conducting linguistic research, it must be emphasized that progress in the field of computer technology entails progress in the creation and improvement of automatic text processing programs and, as result, generates new paradigms of linguistic research.

The authors are aware that some of the information given in the manual will become obsolete over time, that certain specific programs will be replaced by more advanced and multifunctional ones, new corpora will appear, their addresses on the

Internet will change. The authors saw their task in describing a new trend in linguistics as such and the state of corpus linguistics at the beginning of the second decade of the 21st century.

First of all, the authors want to show how, based on corpora, it is possible to work with real language material faster and more efficiently. This book provides examples of research in only a few areas of linguistics - lexicography, grammar and discourse analysis. However, the scope of corpus data in linguistics is much wider. The Appendix contains a brief glossary of corpus linguistics terms.

The authors hope that students of the "Foreign language and literature (English language)", "Linguistics (English language)" directions will be interested in using corpora, regardless of the scope of their scientific interests, and each teacher will find in the manual what his audience needs to talk about.

1. Basic concepts of Corpus Linguistics. Definition of Corpus and Corpus Linguistics

Plan:

1. What is a Corpus?
2. Criteria of Corpus Linguistics
3. Different approaches to corpus definition

The term corpus is derived from the Latin word corpus that means "body". The Latin term, however, displays two distinct descendants in modern English:

a) corpse (it came via Old French cors)

b) corps (it came via modern French corps in the 18th century).

The first form (i.e. corpse) entered into English in the thirteenth century as cors and during the fourteenth century it had its original Latin 'p' reinserted. At first it meant simply 'body', but by the end of the fourteenth century, the sense 'dead body' became firmly established. However, on the other hand, the original Latin term corpus itself was acquired in English in the fourteenth century.

Within the domain of modern corpus linguistics, the term 'corpus' refers to "a large collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language".

Thus, it refers to a large collection of written and spoken text samples, available in machine-readable form, accumulated in scientific manner to represent a particular variety or use of a language. According to scholars, a corpus is a collection of linguistic items that are selected and ordered according to some explicit linguistic criteria defined by the users in order to be used as a sample of a language. It is methodically designed to contain millions of word compiled from diverse text types across many demographic variations to encompass the diversity a natural language exhibits through its multifaceted use. McEnery and Wilson have

classified corpus in a finer scheme of classification characterised by its inherent features:

a) Loosely, a corpus refers to any body of text;

b) Most commonly, it refers to a body of machine-readable text;

c) More strictly, it refers to a finite collection of machine-readable texts sampled to be maximally representative of a language or a variety of it.

In principle, a corpus is actually designed for accurate study of the linguistic properties, features, and phenomena observed in a language. Therefore, we have argued that a systematically compiled corpus, however small in size, should adhere to the following criteria:

• A corpus should faithfully represent both the common and special linguistic features UNESCO – EOLSS SAMPLE CHAPTERS LINGUISTICS - Corpus Linguistics: An Introduction - Niladri Sekhar Dash ©Encyclopedia of Life Support Systems (EOLSS) of the language from which it is designed and developed. The idea of text representation in a corpus indirectly refers to the total sum of its components (i.e. words, phrases, clauses, sentences, etc.) included in it. However, in practice, the total number of words included in a corpus may determine its size but may fail to abide by the principle of proper text representation. Therefore it is better to keep fields open for a corpus as well as keep number of words unlimited for the benefit of language and users.

• A corpus should be large and wide to encompass texts from various disciplines. In other words, directional varieties of language use manifested in various disciplines and domains should have proportional representation in it. For instance, text samples from the fields of natural sciences should carry equal weight as those from aesthetics, literature, mass media, engineering, and social sciences. Thus, a balanced representation of text samples obtained from all disciplines and domains of language use will ensure its reliability.

• A corpus should be a true replica of physical texts available in printed form. Thus, it should faithfully preserve various word forms, spelling variations, punctuation marks as well as various other orthographic symbols used in the

source texts. Else, the actual image of a language or the language variety will be distorted and a corpus will lose its value and authenticity.

• A corpus should be available in the electronic form for easy access by the end users in order to enable common users as well as language researchers to use the database in multiple tasks related to language description and analysis, statistical analysis, language processing, translation, etc. As corpus designers our basic task is to gather large amount of representative text samples covering wide varieties of language used in various domains of our regular linguistic interaction. Since a corpus is capable of representing potentially unlimited selections of text, it may be defined acrostically from the letters used to compose the term in following way:

C : Compatible to both man and computer,

O : Operational in research and application,

R : Representative of a language or a variety,

P : Processable by both man and machine,

U : Unlimited in the amount of data and samples,

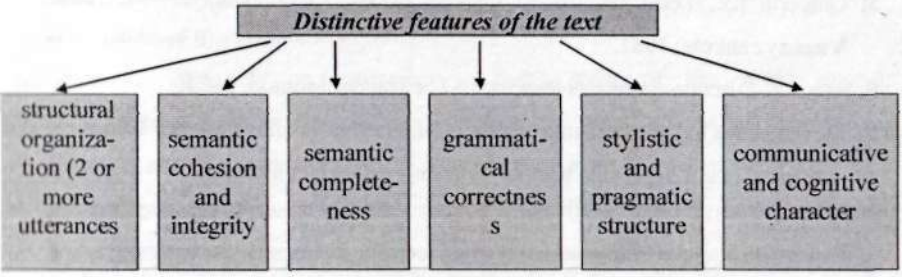
S : Systematic both in formation and representation.

Unless defined otherwise, let us assume that a corpus will possess all the properties mentioned above. Exception may be made for historical corpora, which have limited use due to their diachronic form and composition.

Historical corpora are mostly used within specific areas of historical linguistics that attests indirect importance in the field of empirical language research. In essence, a well-defined and systematically developed corpus is an empirical standard, which acts as a valuable benchmark for validation of usage of all linguistic properties available in a natural language

VISUAL AIDS:

Handout 1. Main characteristics of the texts



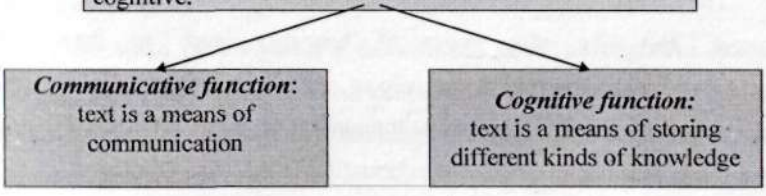
Handout 2

Definition by Galperin I.R.

Galperin I.R. considers that text should be written, only then it has a definite structure, both semantic and grammatical.

Definition by Kolshansky G.V.

Kolshansky G.V.: text is a unit of verbal communication and consists of a number of utterances fulfilling one communicative function.
He stresses two functions of the text: communicative and cognitive.



References

1. Baxtin M. M. Estetika slovesnogo tvorchestva M.: Iskusstvo, 1986.
2. Beaugrande R.A., Dressler W. Introduction to Text Linguistics. –Routledge. 1981.
3. Chernyavskaya V.E. Lingvistika teksta: polikodovost, intertekstualnost, interdiskursivnost. – M.: Librokom, 2009.

4. Dijk T.A. Structures of Discourse and Structures of Power, Communication Yearbook. - № XXII. 1989.
5. Galperin I.R. Tekst kak obyekt lingvisticheskogo issledovaniya. –M.:Nauka. Visshaya shkola, 1981.
6. Harris Z. Discourse analysis Reprints. –The Hague: Mouton, 1963.
7. Moskalskaya O.I. Grammatika teksta. – M.:Nauka. Visshaya shkola, 1981.
8. Turaeva Z.Ya. Lingvistika teksta. – M.: Prosvesheniye, 1986.

2. Types of Corpus

Plan:

1. Princple of representativeness and its different types of packages.
2. Text General Theory
3. *Text Grammar*
4. Text Semantics

The representativeness heuristic is a type of cognitive bias or mental shortcut. Just like other types of heuristics, such as the availability heuristic and anchoring bias, it can help us reduce the time and effort needed to make reasonably good judgments. At the same time, it can lead us astray because we only pay attention to a subset of information: in this case, similarity.

Under the representativeness heuristic, we estimate the probability of something belonging to a specific category based on the degree to which it resembles (or is representative of) the typical or average member of the category. This is called a prototype.

There are a few reasons that can explain why we draw on the representativeness heuristic to make judgments:

- We use mental shortcuts or heuristics to solve problems. Just like with other types of heuristics, our brains are constantly trying to save us time and effort

while navigating a complex world. For this reason, we use simple strategies or rules of thumb to assess information and generate quick responses in everyday tasks. Although heuristics yield fairly good responses, the downside is that they often cause us to oversimplify reality.

- We rely on categories to make sense of the world around us. Categories help us organize and interpret the vast amount of information available in our environment. These categories are built around certain prototypes of what the average member of the category looks like. For example, we recognize a snake when we see one, even if we don't know exactly what type of snake it is. Intuitively, we know we should be careful. By grouping similar things together, we draw on our knowledge of the category and immediately know what to do.
- We overestimate the importance of similarity and ignore more relevant information. Reliance on similarity leads people to ignore "base rate" information, or how often an event occurs. For example, when asked to categorize a person as an engineer or a lawyer, most people would immediately categorize that person as an engineer if they were told that the person enjoyed physics at school—even if they knew that the person was drawn from a population consisting of 90% lawyers and 10% engineers.

In the wake of Langacker's foundational work on cognitive grammar, cognitive linguistics has long been concerned with how speakers represent, process, and actually use language. Although first-generation cognitive linguistics was theory-driven, the usage-based approach of language has been at the core of cognitive linguists' work from the start. In parallel, corpus linguistics has developed as one source of evidence for improving descriptions of the structures and use of languages. It can be described as an approach that empirically analyzes language use in large and principled collections of authentic texts, thanks to automatic and/or computerized tools and based on a combination of quantitative and qualitative techniques. The introduction of corpus-based methods in usage-based cognitive linguistics, also described as its "empirical turn", can be traced

back to the 2000's. Over the last few years, these methods have thrived, relying on increasingly complex statistical tools.

Collecting and analyzing corpus data in usage-based approaches to linguistic investigation relies on one major assumption, namely that the corpus is representative of the linguistic phenomenon under scrutiny. But, of course, corpus representativeness itself is a construct, both a theoretical and a methodological one: language corpora are tools constructed by linguists, and their structural limitations constrain and condition the validity of linguistic findings. Data gathering from a corpus and theorizing are not separate activities: the textual instance is valued as a window on the linguistic system. This is especially true of spoken corpora: any form of transcription operates a drastic selection on the original spoken material, and as such transcription already constitutes a form of theorization.

Multiple theoretical, methodological and practical questions are raised by the issue of corpus representativeness, which, although by no means new, continues to draw attention in current research. A first question is whether corpus representativeness is achievable at all, and if so, how we can identify the relevant criteria to decide whether a corpus represents language use. If corpus representativeness cannot be fully gauged empirically, researchers must remain aware of the extent to which representativeness relies on intuition. Since a corpus cannot realistically be representative of all features of language use, one can wonder how bias in sampling can be addressed. A related issue concerned with corpus design is to what extent representativeness necessarily entails balance. More broadly, the question remains open as to whether the design of a corpus can be totally free from any form of theorization, i.e. whether "pure corpora" can exist. Some answers to these questions have already been provided. Indeed, the representativeness of written corpora may rely on a variety of features, namely variability, sampling and balance.

According to Biber, "representativeness refers to the extent to which a sample includes the full range of variability in a population." Variability can be defined as the interaction between situational (e.g. format, setting, author,

addressee, purposes, topics) and linguistic, distributional parameters (e.g. frequencies of word classes). Sampling is usually based on extralinguistic (sociological, demographic) criteria. Balance, i.e. a proportion of sampled elements that reflects their frequency in the targeted language, is claimed to characterize some corpora and the Lancaster-Oslo-Bergen corpus, though it is not a prerequisite.

Although increasingly larger corpora, including monitor corpora, can be compiled from the Web, large size is not necessarily a priority. As Fillmore put it, "I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way". "Big is beautiful" in the realm of corpora is, perhaps, a "delusion".

Large corpora are often presented as an ideal but, in practice, "small" corpora can go a long way in such domains as gesture studies, English language teaching, the study of metaphors and dialectology, among others. Parallel corpora, i.e. collections of original texts and their translations in one or more languages, are particularly useful in areas of research such as contrastive linguistics, translation studies and computational linguistics, but their alleged lack of representativeness has called for inventive ways of using them.

In the area of spoken corpora, collecting data that represents the variability of the multiple dimensions of speech (phonology and phonetics, prosody, gesture) remains a challenge today. Collecting, transcribing, annotating and analyzing data, is a slow, sometimes complicated, task. Although phonological and prosodic annotations can be partially systematized, technological advances are yet to be made in the automatic recognition of speech and gesture in interactional contexts. Automatic motion capture technologies for gesture research are promising, but so far little advanced. As part of initiatives such as the TGIR Huma-Num Multi-

Com–CORLI Consortium, multimodality researchers collaborate to develop collective harmonized practices of collection, transcription and archiving of spoken corpora. Overall, even if many advances are yet to be made in the construction of representative spoken corpora, the field is making fast progress.

Given the multiplicity of issues raised by corpus representativeness and the limited space of a journal's special issue, the contributions assembled here are not intended to give a comprehensive overview of the topic in linguistics today. Rather, they allow to address a number of specific issues raised across various approaches and types of data and in relation to different research goals within linguistic traditions grounded in or having close theoretical and methodological links with cognitive linguistics. All the contributions stem from presentations given at the AFLiCo JET workshop "Corpora and representativeness", which was held on May 3-4, 2018 at the Université Paris Nanterre.

One major issue is the articulation between intuitive and corpus-based approaches to language study. Traditionally, usage-based models have sought to ground linguistic analysis in the observation of naturally occurring data, which, they argue, offers a more representative picture of language than does introspective evidence. Interestingly, the habitual dichotomy between the two approaches is softened at several points in the thematic issue. Corpus methodologies are not necessarily meant to eliminate intuitions altogether but rather to provide support for them and, from there, for the construction of linguistic theories. For the corpus linguist, intuition may prove to be particularly useful in detecting corpus oddities. Rocking the usage-based linguistics boat, Newmeyer goes as far as to claim that data drawn from big corpora yield essentially the same results as intuition-based data.

The fundamental rationale for relying on corpora is the unique possibility they offer of using quantitative data in linguistic analysis. In usage-based linguistics, frequencies of occurrences and co-occurrences of forms are key to elaborating theoretical models. This is not to say that qualitative analyses should be abandoned. In this special issue, linguists working in various theoretical

frameworks stress the importance of integrating quantitative and qualitative approaches. While the necessary back-and-forth methodological movement between frequencies and specific contexts of use is either explicitly addressed or exemplified in practice throughout the usage-based contributions, approaches not traditionally based on usage and/or data quantification – enunciative theory, generative grammar – seize the opportunity to use quantitative tools to complement qualitative research.

A second issue addressed in the present collection of articles is corpus size. In specific avenues of research, notably the study of grammatical structures on the basis of conversational corpora, "big" can be argued to be indeed "beautiful". But a number of disadvantages can also be pointed out. First, as mentioned by Lemmens, without a fair amount of knowledge of sophisticated statistical tools, it is difficult, if not impossible, to make sense of large volumes of data, so that big or "mega" corpora in the end may turn against the researcher. In other cases, big corpora may simply be no more efficient than corpora of more reasonable size, since there is a point at which the saturation of new information is reached during corpus construction.

Finally, a more crucial aspect than size, as it emerges from a number of articles, is sampling. A corpus may be small but more representative of a language, variety or register than larger ones if sampling is based on systematic, linguistically-motivated decisions rather than convenience or some principle of authority, as was perhaps often the case with first-generation corpora. In the debate over whether sampling should aim at representing the full range of texts produced (i.e. production-based sampling) or the diversity of high-impact texts (i.e. reception-based sampling), the authors addressing the question explicitly or implicitly advocate production, each stressing a different reason. Sampling according to production rather than reception acts as a safeguard against including in the corpus a lot of intentionally unusual material which is recognized as such by the language user and has arguably no influence on their productive linguistic system. Only a production-based sampling method enables the corpus compiler to

gauge and represent the diversity of a given genre or register. Specifically, using the production criterion ensures that the corpus does not over-represent features of individual speakers at the expense of others.

References

1. Beaugrande R.A., Dressler W. Introduction to Text Linguistics. –Routledge. 1981.
2. Chernyavskaya V.E. Lingvistika teksta: polikodovost, intertekstualnost, interdiskursivnost. – M.: Librokom, 2009.
3. Dijk T.A. Some Aspects of Text Grammar. –The Hague-Paris. 1972.
4. Galperin I. R. Stylistics. –M.: Higher school, 1977.
5. Galperin I.R. Tekst kak obyekt lingvisticheskogo issledovaniya. –M.:Nauka. Visshaya shkola, 1981.
6. Halliday M. and Hasan R. Cohesion in English. –London: Longman, 1976.
7. Moskalskaya O.I. Grammatika teksta. – M.:Nauka. Visshaya shkola, 1981.
8. Turaeva Z.Ya. Lingvistika teksta. – M.: Prosvesheniye, 1986.

3. Technology of Corpus Linguistics

Plan:

1. Corpus preprocessing construction.
2. Problems of Text Stylistics
3. Problems of Text Linguistics
4. Text type genre

There runs throughout this special issue the tacit agreement that complete representativeness of a language might never be achieved. As a consequence, perhaps, most contributors choose to build or carry out their analyses on corpora focused on some specific type of language production, for which corpus samples may represent the target population more exhaustively: sign language, language

acquisition, political discourse, spontaneous conversation, fictional spoken language and performance texts. Even then, representing those types of language use is not without its challenges. Issues specifically related to spoken and signed language corpora are discussed in a number of contributions. Given the complexity of this type of unscripted data, one may wonder to what extent it may serve as a basis for the study of grammar, for instance. In this special issue, two different ways of handling such complexity are presented. The first one is to turn it into an object of study in itself and seek solutions for how to best represent this complexity. In the field of sign languages, this means creating the most accurate tools for the transcription, visualization and searchability of data.

The second one is to develop an adaptation strategy and rely on data drawn from fictional spoken language and performance texts as the closest representation of speech. In terms of corpus construction and use, this nonetheless involves two major challenges, as the corpus should be representative of the performance genre itself, which in turn, should be representative of the spontaneous spoken interaction the genre is taken to reproduce, despite its inherent planned and editable nature. Arguably, although fictional spoken language and performance texts are an approximation of on-line speech production, the gap between the representation and the reality represented may be significantly reduced through adequate sampling techniques.

Although the many echoes between the various contributions suggested more than one way of structuring the special issue, we have decided to group them under three sections:

I. Representativeness in (target-specific) corpus construction

II. Addressing representativeness through case studies

III. Theorizing corpus representativeness

The first section, dedicated to representativeness in (target-specific) corpus construction, includes three papers written by Boutet et al., Parisse and Perrez et al. respectively. The first contribution by Dominique Boutet et al. presents the typefont Typannot, a transcription system dedicated to the annotation of sign

SamDCHTI ARM

languages. Unlike vocal languages, which are monolinear in nature, sign languages are characterized by multilinearity, since meaning is expressed simultaneously with various distinct articulators (e.g. hands, face, body). Other existing annotation systems remain limited: they either fail to provide information on form altogether, are cumbersome when describing language signs on the basis of form, or fail to account for the multilinearity of sign languages. Representing sign language in graphic form is hence a major challenge for sign language corpus developers. With Typannot, Boutet et al. take up that challenge. This modular typographic system aims to integrate three levels of information: the parameter (handshape, orientation, location, movement, facial expression), the components of the parameter, and the characteristics of each component, so as to provide a fine-grained description of the form of language signs for optimal representativeness of the data. Typannot's typefaces are created on the basis of the four underlying principles of scriptability, readability genericity, and modularity: they are meant to be easy to write and read, as well as to account for both low-level and concatenated information. Typannot stems from an interdisciplinary research involving linguists, designers and developers, and aims at being transferable to all existing sign languages.

Christophe Parisse's paper contributes to research on dense corpora by assessing the optimal size of a longitudinal dense corpus. As he explains, dense longitudinal corpora aim to be as representative as possible of a child's language development. They allow tracing the development of linguistic knowledge based on numerous samples of child's speech and input over relatively short periods of time. Since the compilation of dense corpora is time-consuming, identifying the optimal minimum size of this type of corpus is crucial. To do so, Parisse relies on two measures: word-metric, used to represent the development of lexical knowledge, and bigram-metric, used to represent the development of basic syntactic knowledge over time. Measuring word-metric and bigram-metric allows him to determine the smallest number of sessions to be included in a longitudinal dense corpus, in order to provide enough data to predict the use of a word or a

bigram. His results show that the quality of both word and bigram coverage obtained in about 40 sessions is high enough to study language acquisition. Of course, this result does not undermine the value of very large dense corpora, but shows that high-quality research can be conducted on smaller dense corpora. His paper therefore sets a useful practical upper limit in the area of dense corpus construction.

The question at the heart of Julien Perrez et al.'s article is the following: how is the category of political discourse (or should be) defined? How homogeneous is it? Linguistics studies of political discourse bear mostly on productions by political elites such as presidential or parliamentary debates, presidential addresses or public speeches, and tend to leave aside other forms such as media or citizen discourses. Their article aims to define the genre of political discourse, based on both extralinguistic and linguistic features. First, they apply the bibliometric method PRISMA (borrowed from the political sciences) on a sample of 172 scientific articles from the Scopus database, so as to map out what types of discourse have been categorized as political in linguistic research over the past twenty years, and to identify their extralinguistic features (e.g. type of actors, materials, themes, geographical origin). Second, they assess how consistent the notion of political discourse is from a linguistic point of view. Using multidimensional register analysis, the authors study the formal linguistic features of three subtypes of political discourse (parliamentary debates, televised debates and citizen corpora) so as to assess similarities between their textual registers. Results show great divergence, suggesting that political discourse is best defined as an abstract generic category for a range of different registers whose linguistic features are sensitive to the situational context of use.

The second section of this special issue addresses representativeness through three case studies proposed by Grieve-Smith, Terry and Ranger, respectively. In the first of these contributions, Angus Grieve-Smith, driven by the challenge of how to best represent the way people used to talk in 19th-century France, decides to turn to the language of the theater as one of the closest genres to spoken

conversation. A corpus of theatrical productions can be considered to be representative of spoken language, he argues, providing adequate sampling methods are used. Grieve-Smith shows that the 'principle of authority' behind the sampling of the drama section of FRANTEXT results in the under-representation of some authors and a general bias towards formal language features. The author then presents his ongoing compilation project, the Digital Parisian Stage corpus, for which he uses as a sample frame an exhaustive list of every play that premièred in Paris in the nineteenth century. The representativeness of FRANTEXT and the Digital Parisian Stage corpus are then compared through a case study of the various syntactic realizations of negation. The results show that the Digital Parisian Stage corpus provides a more accurate picture of the language used over the period, including recent grammatical innovations.

Starting from the general assumption that the language of TV series is a polished, but truthful version of naturally occurring conversation, Adeline Terry's paper deals with the following research questions: are the metaphors in TV series representative of those that can be found in naturally occurring conversation? Are the main source domains used similar in naturally occurring conversation and in TV series? To answer them, she compares results drawn from a corpus of TV series and results of previous studies on metaphor in non-fictional corpora. Major differences stand out: compared to naturally occurring conversation, metaphors in the TV series corpus are more often creative and/or extended, are more used for humorous functions, and are used for characterization, dramatization, and aesthetic purposes. Terry also highlights the multifunctionality of metaphors in the TV series corpus. Differences in the use of metaphors between real-life conversation and TV series discourse suggest that they constitute distinct genres, and hence that TV series discourse is a genre that should be studied for itself, and not as a representative of real-life conversational discourse.

In the last contribution of the section, Graham Ranger brings together the methodological strengths of corpus linguistics with the robust theoretical tools of the French Théorie des Opérations Prédicatives et Énonciatives in a case study of

the English marker 'along'. The author starts with a critical appraisal of the status of linguistic data in the enunciative framework. Often drawn from a single genre, realistic fiction, linguistic evidence is traditionally submitted to manipulations about which intuitive acceptability judgements are made. According to Ranger, this methodology goes against the enunciative approach's commitment to study language in the diversity of its naturally occurring manifestations. The author then proposes a corpus study of 'along' using the data of the British National Corpus. Using statistical measures of occurrences and co-occurrences of terms, he identifies four contextual configurations, which are mapped with four main values – spatial, temporal, subjective and argumentative – for the marker. In each value, 'along' marks a dynamic identification between a locatum and a locator, construed as an unbounded, sequentially ordered space. The specificity of the marker is highlighted through a comparison with the compound form 'alongside', a closely related but semantically distinct marker, as evidenced by corpus data.

The third and final section of the special issue focuses on theorizing corpus representativeness. It includes a paper by Egan, and a tandem paper by Newmeyer and Lemmens. Thomas Egan's contribution addresses some practical and theoretical issues of representativeness in the design and use of corpora. The author discusses the ideal composition of a written corpus, which, he argues, should be based on production and avoid text types likely to misrepresent actual language use (e.g. grammar books, poetry, historical fiction). He then turns to multilingual corpora, for which he presents methods to obtain maximally representative data and avoid translation effects. Specifically, he shows the benefits of two models: the (expandable) four-text model and the three-text model. In the four-text model, the researcher makes use of bidirectional corpora, which contain original texts and translations in each of the languages represented. In the three-text model, linguistic analysis is based on the comparison of productions in two target languages, while the source language functions as tertia comparationis.

In the last contribution, Frederick Newmeyer and Maarten Lemmens use the original format of a tandem paper to bring into discussion generative grammar and

usage-based cognitive linguistics, two traditionally contrasted theoretical frameworks, on the issue of corpus data and representativeness. The article starts with Newmeyer's demonstration of the limits of small conversational corpora when used to build theoretical models of grammar. The author goes on to reaffirm the relevance of introspective data in grammatical theory, since, he argues, they provide essentially the same insight into grammatical structure than data obtained from big corpora. Newmeyer's conclusions prompt Lemmens to enlarge the perspective and engage in a defence of usage-based linguistics over intuition-based approaches, highlighting the "complex and all-pervasive role of frequency with respect to the structure of grammar" and using the opportunity to give an updated overview of the field. The arguments proposed by the two authors in the initial position and response statements give rise to genuine discussion in Newmeyer's rejoinder and Lemmens's final acknowledgment. Closer connections than usually assumed are suggested between the two theoretical approaches, in particular with respect to their underlying assumptions and goals.

In stylistics, 'text' means a piece of information that can be spoken or written appearing in its original form. It refers to language use that has not been altered by either summarizing or paraphrasing. Stylistics involves analysis of rhetoric, tone, irony, metaphors, and such styles applied in the use of language. In stylistics, text can be intricate, which means precise information. Such texts usually require keen analysis to interpret because they are complicated. Surface texts are words whose meaning is as simple as they come out. The texts are easy to interpret because what the listener hears is precisely what the speaker means. Literary texts involve the unrestricted use of languages such as poetry and stories. They are mainly used for entertainment purposes and may involve fictional information.

Text linguistics is a branch of linguistics that deals with texts as communication systems. Its original aims lay in uncovering and describing text grammars. The application of text linguistics has, however, evolved from this approach to a point in which text is viewed in much broader terms that go beyond a mere extension of traditional grammar towards an entire text. Text

linguistics takes into account the form of a text, but also its setting, i. e. the way in which it is situated in an interactional, communicative context. Both the author of a (written or spoken) text as well as its addressee are taken into consideration in their respective (social and/or institutional) roles in the specific communicative context. In general it is an application of discourse analysis at the much broader level of text, rather than just a sentence or word.

"A text is an extended structure of syntactic units [i. e. text as super-sentence] such as words, groups, and clauses and textual units that is marked by both coherence among the elements and completion ... [Whereas] a non-text consists of random sequences of linguistic units such as sentences, paragraphs, or sections in any temporal and/or spatial extension."

"A naturally occurring manifestation of language, i. e. as a communicative language event in a context. The *surface text* is the set of expressions actually used; these expressions make some knowledge *explicit*, while other knowledge remains *implicit*, though still applied during processing."

"A term used in linguistics to refer to any passage- spoken or written, of whatever length, that does form a unified whole [....] A text is a unit of language in use. It is not a grammatical unit, like a clause or a sentence; and it is not defined by its size [....] A text is best regarded as a *semantic* unit; a unit not of form but of meaning."

"A text is made up of sentences, but there exist separate principles of text-construction, beyond the rules for making sentences."

"Text is a set of mutually relevant communicative functions, structured in such a way as to achieve an overall rhetorical purpose."