# The Bulgarian National Corpus: Theory and Practice in Corpus Design

*Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova*
Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria

## ABSTRACT

The paper discusses several key concepts related to the development of corpora and reconsiders them in light of recent developments in NLP. On the basis of an overview of present-day corpora, we conclude that the dominant practices of corpus design do not utilise the technologies adequately and, as a result, fail to meet the demands of corpus linguistics, computational lexicology and computational linguistics alike.

We proceed to lay out a data-driven approach to corpus design, which integrates the best practices of traditional corpus linguistics with the potential of the latest technologies allowing fast collection, automatic metadata description and annotation of large amounts of data. Thus, the gist of the approach we propose is that corpus design should be centred on amassing large amounts of mono- and multilingual texts and on providing them with a detailed metadata description and high-quality multi-level annotation.

We go on to illustrate this concept with a description of the compilation, structuring, documentation, and annotation of the Bulgarian National Corpus (BulNC). At present it consists of a Bulgarian part of 979.6 million words, constituting the corpus kernel, and 33 Bulgarian-X language corpora, totalling 972.3 million words, 1.95 billion words ($1.95 \times 10^9$) altogether. The BulNC is supplied with a comprehensive metadata description, which allows us to organise the texts according to different principles. The Bulgarian part of the BulNC is automatically processed (tokenised and sentence split) and annotated

*Keywords: corpus design, Bulgarian National Corpus, computational linguistics*

at several levels: morphosyntactic tagging, lemmatisation, word-sense annotation, annotation of noun phrases and named entities. Some levels of annotation are also applied to the Bulgarian-English parallel corpus with the prospect of expanding multilingual annotation both in terms of linguistic levels and the number of languages for which it is available. We conclude with a brief evaluation of the quality of the corpus and an outline of its applications in NLP and linguistic research.

## 1 INTRODUCTION

Since the first structured electronic corpus, the Brown Corpus (Francis and Kučera, 1964), corpora have been increasingly used as a source of authentic linguistic data for theoretical and applied research. Corpus-based studies have been employed in various areas of linguistics, such as lexicology, lexicography, grammar, stylistics, sociolinguistics, as well as in diachronic and contrastive studies (Meyer, 2002).

Traditional definitions of a corpus emphasise different aspects. A corpus is typically viewed as a collection of authentic linguistic data that may be used in linguistic research (Garside *et al.*, 1997). Sinclair (2005) adds to this definition the storage format and the selection criteria: "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.*" Finally, annotation at different linguistic levels (phonological, lexical, morphological, morphosyntactic, syntactic, semantic, discourse and stylistic) amplifies the corpus's value by extending its functionalities and applications (McEnery *et al.*, 2006). One of many different definitions states: *A corpus is a large collection of language samples, suitable for computer processing and selected according to specific (linguistic) criteria, so that it represents an adequate language model.* (Koeva, 2010).

With the increased development of language technologies, the applications of corpora have been extended to all areas of computational linguistics and natural language processing (NLP). Corpora have become an indispensable resource for generating training sets for machine learning, language modelling, and machine translation. These developments have led to the necessity for reconsidering the traditional notions in corpus linguistics. As a result, we propose a corpus design based on automatic collection of very large monolingual and

multilingual (and in particular parallel) corpora that cover a wide variety of styles, thematic domains, and genres.

This paper contributes to the discussion on the perspectives of corpus development in three ways: (i) by reconsidering several key traditional principles underlying corpus design, (ii) by proposing an approach in corpus design based on the revision of those fundamentals in light of recent advances in NLP technologies, (iii) by illustrating how the proposed model is applied in the Bulgarian National Corpus (BulNC).

The study is placed in the context of well-known corpora, both mono- and multilingual (Section 2), with an outline of their general features. The concepts of corpus size, balance, and representativeness are discussed in Section 3. In the same section we present our concept of corpora, which integrates the best practices of traditional corpus linguistics with the potential of the latest technologies for web crawling and language processing. Section 4 presents the process of compiling, structuring, documenting, and annotating the BulNC, followed by a brief evaluation of the quality of the corpus and an outline of some current applications.

## 2 OVERVIEW OF CONTEMPORARY MONOLINGUAL AND MULTILINGUAL CORPORA

The last decades have seen the compilation of large mono- and multilingual corpora for a lot of languages, including some less-resourced ones, Bulgarian among them. The brief overview illustrates the current standards in corpus design and compilation and provides a point of departure for comparison with the proposed paradigm.

### 2.1 *Large monolingual corpora*

1. At the time of its creation, the British National Corpus[1] (BNC) was one of the biggest (100 million words) existing corpora. Being compiled according to carefully devised principles and classification criteria[2], it set the standards for general monolingual synchronic corpora for quite some time. The BNC represents not

[1] `http://www.natcorp.ox.ac.uk`

[2] `http://www.natcorp.ox.ac.uk/corpus/creating.xml`

only written, but also spoken language, respectively 90% and 10% of the samples. It is POS-tagged, lemmatised, and supplied with detailed metatextual information. The corpus (text and annotated data) can be searched both online – through various search tools, and offline using XAIRA[3].

2. The Corpus of Contemporary American English[4] (COCA) is a 450+ million-word corpus currently in progress with an increase rate of 20 million words per year. The texts are evenly divided between 5 categories – spoken language, fiction, popular magazines, newspapers, and academic writing (Davies, 2010), each category currently containing 90 to 95 million tokens (as of June 2012). The corpus provides a web search interface (shared with the Google Books corpora) that allows searches for regular expressions and specifications for POS, lemma, collocations, frequency and distribution of synonyms. The queries may be refined in terms of genre or time period.
3. The Slovak National Corpus[5] (SNK) contains more than 719 million tokens[6]. The texts are divided into several categories with the following distribution: journalism (73%), literary texts (14%), professional texts (12%), and other (1%). A subcorpus of 1.2 million tokens, manually annotated with morphological tags, has also been compiled. The SNK and its subcorpora can be searched with a CQL (Corpus Query Language) compatible query syntax (Christ and Schulze, 1994) through a web interface or via the Bonito client[7], cf. the Czech National Corpus.
4. The Croatian National Corpus[8] (HNK) includes about 101 million words of mainly contemporary Croatian texts that cover different media, genres, styles, fields, and topics. They fall into the categories of informative texts (74%), fiction (23%), and mixed texts (3%), with further subdivision within these categories. The morphological tagset used in the HNK annotation is Multext-East-

[3] `http://www.natcorp.ox.ac.uk/tools/index.xml`
[4] `http://corpus.byu.edu/coca/`
[5] `http://korpus.sk`
[6] The version released at the beginning of 2011
[7] `http://korpus.juls.savba.sk/usage_en.html`
[8] `http://www.hnk.ffzg.hr/cnc.htm`

compatible, and the corpus can be searched offline through the Manatee/Bonito server-client[9].

5. The Russian National Corpus[10] (RNC) comprises more than 300 million words of texts ranging from the middle of the 18th century to the present day. The main part of the corpus, about 100 million words, consists of contemporary texts of three general categories: fiction (40%), non-fiction (56%) and recordings of public and spontaneous speech (4%), with a detailed internal classification[11]. The corpus has been automatically supplied with morphosyntactic annotation, and parts of it have been manually verified and disambiguated. A portion of the corpus has also been annotated with syntactic dependencies and semantic roles. Lexical-semantic information, covering taxonomic, mereological, topological and other features of words, has been assigned. The RNC provides a web interface for detailed search in the whole corpus and its subcorpora for words and phrases, grammatical (POS, morphology), syntactic, and semantic (taxonomy, evaluation and mereology) features.

6. The Czech National Corpus[12] (CNC) (Kocek *et al.*, 2000) was started in the 1990s. Since then it has been constantly growing and according to the latest published estimates currently amounts to 1.3 billion words. It consists of a number of subcorpora, among them several balanced subcorpora of 100 million words each, compiled every several years, the latest version being SYN2010. The distribution of texts across categories is as follows: fiction (40%), technical literature (27%) and journalism (33%), with more elaborate subdivision within these categories. Most of the written corpora in the CNC are annotated. The CNC can be searched for words and phrases using exact match and regular expressions both online and offline through the Corpus manager Manatee and the client Bonito[13].

[9] `http://www.hnk.ffzg.hr/pretraga_en.html`
[10] `http://www.ruscorpora.ru/`
[11] `http://www.ruscorpora.ru/en/corpora-stat.html`
[12] `http://ucnk.ff.cuni.cz`
[13] `http://www.textforge.cz/download`

7. The National Corpus of Polish[14] (NCP) (Bański and Przepiórkowski, 2010) contains fiction, daily newspapers, specialised periodicals and journals, transcripts of conversations and Internet texts, amounting to approximately 1 billion words. A balanced 250-million-word subcorpus extracted from the NCP has been compiled (Przepiórkowski, 2011), and part of the NCP, approximately 1.2 million words, was manually annotated. The corpus can be searched online through two search engines (Poliqarp and PELCRA) that allow queries for words and regular expressions; the former also searches for morphological tags and the latter offers collocation extraction. The search may be further refined with editorial and descriptive (genre or domain) metadata.
8. The German Reference Corpus[15] (DeReKo) amounts to 5.4 billion words (Bański *et al.*, 2012).The concept of the corpus explicitly rejects the feasibility of balance and representativeness (Kupietz *et al.*, 2010). The corpus is conceived as a versatile "primordial" sample from which specialised subsamples, or "virtual" corpora, are drawn. The development of the corpus is focused on the maximisation of size and stratification, rather than on the composition of specialised subsamples. The corpus includes POS annotation, partial morphological disambiguation, named entities, and syntactic dependencies. The corpus can be searched offline through the COSMAS II client[16].
9. A number of large corpora have recently come into existence, with size ranging from several (Baroni and Kilgarriff, 2006; Pomikálek *et al.*, 2009), through dozens (Pomikálek *et al.*, 2012), to hundreds of billions of words (Google Books Corpora, GBC[17], the largest being the 200-billion-word GBC of American English). What distinguishes these from the rest of the discussed corpora is that they represent a different type of approach to corpus creation, since they are collected fully automatically from web content. The GBC web search interface allows queries according to several criteria:

[14] `http://nkjp.pl`

[15] the Archive of General Reference Corpora of Contemporary Written German, `http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html`

[16] `http://www.ids-mannheim.de/cosmas2/`

[17] `http://googlebooks.byu.edu/`

exact words or phrases, regular expressions, POS, lemma, collocations, frequency and distribution of synonyms, with further refinement in terms of genre or time period.

This brief outline shows that the dominant and constant tendency is for corpora to aim at a size ranging from several hundred million up to over a billion words.

All corpora are at least partly annotated and provide online or offline search interfaces. The differences lie in the quantity of the annotated data and the levels of annotation. The minimum annotation is generally POS tagging. Many corpora also include morphosyntactic annotation and lemmatisation, and some provide syntactic (e.g., DeReKo) or semantic annotation (e.g., COCA, GBC).

Some of the corpora discussed here follow predefined structure and classifications, whereas others abandon balance and representativeness in favour of size. The design criteria differ not only when it comes to coverage and distinction of textual categories and subcategories, but, more fundamentally, in the underlying assumptions. Balance is viewed as the equal representation of predefined text categories (Davies, 2010), or as a distribution of texts proportional to language production (Atkins *et al.*, 1991) or language reception estimated according to various criteria. Some authors, involved in the compilation of the previously discussed corpora, have proposed assessments of language reception on the basis of stylistic (Przepiórkowski *et al.*, 2010), sociological (Čermak and Schmiedtová, 2003), and marketing (Tadić, 2002) surveys.

The following trends emerge with respect to the relationship between size, balance and representativeness:

- Creation of corpora according to a predefined methodology that is considered sufficiently adequate to ensure corpus balance and representativeness (1-5).
- Development of large unbalanced corpora paired with static balanced subcorpora that are compiled in accordance with a carefully devised structure (6-7).
- Compilation of large unbalanced corpora that enables the extraction of subcorpora based on metadata description (8).

- Compilation of very large unbalanced corpora from the web whose structure and content are not concerned with balance and representativeness (9).

### 2.2 *Large parallel corpora*

1. Some of the major parallel corpora that have been largely drawn on by the NLP community are multilingual repositories of publicly available legal, administrative or journalistic texts, such as: the European Parliament Proceedings Parallel Corpus[18] (Europarl), the Canadian Hansard Corpus of parliamentary proceedings; the News Commentary Corpus[19]; the JRC-Acquis Multilingual Parallel Corpus[20] of legal texts, the EU Official Journal[21], MultiUN[22]. The OPUS collection[23] includes a set of various corpora – administrative (e.g., the EMEA corpus of administrative medical texts), news (including the SETimes corpus of news in eight Balkan languages and English), etc.
   These corpora are distinguished from traditional ones in that the data have been compiled for a different purpose and have subsequently been employed as corpora. Therefore not all of them are annotated. OPUS, EuroParl, and JRC-Acquis are tokenised, sentence-segmented and sentence-aligned. Parts of the OPUS collection are POS-tagged for some languages, with word alignment currently under way and dependency parsing envisaged in the near future. A part of the Hansard corpus is also sentence-split and aligned.
2. The Czech-English parallel corpus (CzEng) comprises 206.4 million tokens in Czech and 232.7 million tokens in English, distributed across 7 source domains: fiction, EU legislation, movie subtitles, parallel webpages, technical documentation, news, and texts from Project Navajo[24]. The predominant domains are fiction and legislation. The texts have gone through automatic sentence-

[18] http://www.statmt.org/europarl
[19] http://www.statmt.org/wmt11/translation-task.html
[20] http://langtech.jrc.it/JRC-Acquis.html
[21] http://eur-lex.europa.eu/en/index.htm
[22] http://www.euromatrixplus.net/multi-un/
[23] http://opus.lingfil.uu.se/
[24] http://ufal.mff.cuni.cz/czeng/czeng10/

splitting and alignment. Morphosyntactic tagging, lemmatisation, word alignment, surface and deep-level syntactic annotation are provided (Bojar *et al.*, 2012).

3. The Hunglish corpus[25] is a sentence-aligned parallel corpus of Hungarian and English containing 34.6 million Hungarian and 44.6 million English words. The texts cover a number of varied domains: literature, religion, international law, movie subtitles, software documentation, magazines, and business reports (Varga *et al.*, 2005). The corpus has been tokenised with the rule-based HunToken tokeniser and stemmed with the Hunmorph morphological analyser.
4. The Polish-Russian Parallel Corpus[26] consists of 50 million words equally divided between Polish originals with their Russian translations and the other way round. The corpus includes classical and modern literature as well as legal and journalistic texts. The texts are annotated according to the annotation schemes of the National Corpus of Polish and the Russian National Corpus.
5. The Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[27] is a corpus of 440,000 parallel sentences made up of 12 million Japanese words and 11.5 million English words. The texts concerning Kyoto and other specific topics, such as traditional Japanese culture, religion, and history, are manually translated into English, aligned and verified. The corpus was used for the development and evaluation of Japanese-English machine translation systems in the Kyoto Free Translation Task (Neubig, 2011).

Many of the available parallel corpora are of modest size, especially in comparison with monolingual corpora, and as a rule they belong to a limited number of domains determined by the availability of parallel texts. For the most part these corpora are compiled automatically by web crawling or by downloading publicly available parallel collections. More varied content can be obtained from publishers or through manual compilation, but these methods are less efficient. A third source of parallel data has been the translation of monolingual corpora; Xu and Sun (2011), among others, have experimented with

[25] `http://mokk.bme.hu/resources/hunglishcorpus/`

[26] `http://www.pol-ros.polon.uw.edu.pl/`

[27] `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`

machine translation for increasing parallel data for less common languages.

Due to the limited domain and genre diversity of parallel texts, balance and representativeness are usually considered irrelevant. Three dominant patterns of parallel corpus design emerge:

- Acquisition of corpora from repositories of parallel content (1). Being publicly available, these collections are often reused in other corpora either as raw text or with the supplied annotation.
- Compilation (preferably automatic) of parallel corpora that aims at reflecting the diversity of monolingual corpora, possibly using readily available corpora (2-4).
- Construction of parallel corpora by means of human translation (5) or machine translation of the original content.

To conclude, there is considerable heterogeneity among the existing monolingual and parallel corpora in terms of size, design criteria, annotation principles, etc. At the same time, neither the possibilities of the modern technologies, nor the enormous amount of available data are used rationally enough to serve the needs of NLP. Moreover, traditional services for extraction of concordances and collocations fail to meet the needs of modern corpus linguistics and computational lexicography. The automatic collocation dictionaries extracted through "sketch grammars" and an algorithm for finding "good dictionary examples" allow a more efficient access to corpus data (Kilgarriff *et al.*, 2009).

### 2.3 *An overview of Bulgarian corpora*

The work on corpora for Bulgarian began in the 1990s with the compilation of relatively small text collections for specific purposes.

1. Two corpora of Spoken Bulgarian have been created in the 1990s[28] (Nikolova, 1987; Aleksova, 2000).
2. Further efforts were focused on building large reference corpora such as the BulTreeBank Text Archive (since 2000) with 15% of the texts coming from fiction, 78% from newspapers and about 7% excerpted from legal and government texts and others (Simov

[28] `http://folk.uio.no/kjetilrh/bulg/Aleksova/index.html` and `http://folk.uio.no/kjetilrh/bulg/Nikolova/`

*et al.*, 2002). This corpus recently evolved into the Bulgarian National Reference Corpus[29]. The corpus interface executes queries allowing exact matches or regular expressions.

3. The "Brown" Corpus of Bulgarian[30] (BCB) (Koeva *et al.*, 2006), compiled in the period 2001 to 2005 as a general corpus of contemporary Bulgarian, is one of the Bulgarian corpora that closely follow a clearly established methodology, namely that of the Brown Corpus of Standard American English (Francis and Kučera, 1964). Text samples can be searched using queries for exact matches or regular expressions. The "Brown" Corpus of Bulgarian with full-length texts[31] (FullBrown), consisting of the originals from which the BCB 2000-word excerpts were sampled, is included as an integral part of the Bulgarian National Corpus and may be searched through its web interface.
   Concomitantly, a number of Bulgarian annotated corpora have been developed, covering POS tagging, word sense annotation, annotation of dependency structure, and sentence and clause alignment.
4. The Bulgarian POS-Tagged Corpus (BulPosCor) totals 174,697 words, each of them manually annotated with the context-relevant POS and morphosyntactic features.
5. The Bulgarian Sense-Annotated Corpus (BulSemCor) amounts to 95,119 lexical items, covering both single words and multiword expressions. Each of them has been POS-tagged, lemmatised, and assigned a synonym set from the Bulgarian Wordnet (Koeva *et al.*, 2006) that best corresponds to the sense of the lexical item in the particular context.
   Both BulPosCor and BulSemCor are integrated into the BulNC and can be accessed through its web search interface (see Section 4.6.1), as well as through specially developed interfaces[32].
6. The BulTreeBank is a syntactically annotated corpus developed within the HPSG framework (Simov and Osenova, 2004). Access to the data may be gained by submitting a contact form. The

[29] `http://www.webclark.org`
[30] `http://dcl.bas.bg/Corpus/home_en.html`
[31] `http://dcl.bas.bg/en/corpora_en.html`
[32] `http://dcl.bas.bg/poscor/en/`, `http://dcl.bas.bg/semcor/en/`

HPSG-based annotation has later been converted into syntactic dependency annotation covering 214,000 tokens, or slightly more than 15,000 sentences (Chanev *et al.*, 2006).

7. The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) (Koeva *et al.*, 2012a), a parallel sample from the Bulgarian National Corpus, comprises 176,397 tokens for Bulgarian and 190,468 for English. It is supplied with syntactic annotation for sentence and clause boundaries, relations between syntactically linked clauses (i.e., coordination or subordination) and clause-introducing words and phrases.

The main purpose of corpora 4 to 7 is to serve as training and test corpora in the development of various automatic annotation tools for multi-level annotation with sufficient accuracy and coverage. Most of the parallel corpora involving Bulgarian are purpose-driven and cover specific domains, such as administrative texts or fiction, which are widely available in parallel versions and hence easily collected.

The Bulgarian subcorpus in the JRC-Acquis corpus of EU legislation documents contains 16.1 million tokens (Steinberger *et al.*, 2006). The Bulgarian part of the SEE-ERA.NET Administrative Corpus (SEnAC) consists of excerpts from the Acquis communautaire, about 1.5 million tokens and 60,389 translation units, each containing one sentence translated into 8 languages (Tufiş *et al.*, 2009). The EuroParl Corpus of proceedings of the European Parliament includes approximately 6 million tokens in Bulgarian (Koehn, 2005).

The Multext-East corpus incorporates the original and translations into six languages of George Orwell's novel *Nineteen Eighty-Four*, with the Bulgarian part amounting to 54,823 tokens (Dimitrova *et al.*, 1998). In the SEE-ERA.net Fiction Corpus (SEnFC), consisting of translations of Jules Verne's novel *Around the World in 80 Days* into sixteen languages, the Bulgarian part adds up to 58,678 tokens (Tufiş *et al.*, 2009). The Cultural Greek-Bulgarian Corpus is a bilingual collection of literary and folklore texts containing approximately 350,000 tokens (Giouli *et al.*, 2009). The extended RuN-Euro Corpus includes a small Bulgarian part of 366,329 tokens (Grønn and Marijanovic, 2010), and ParaSol (von Waldenfels, 2006, 2011), a corpus of fiction texts, includes a Bulgarian subcorpus of over 2 million tokens as of June 2011. The Bulgarian-Polish-Lithuanian Parallel Corpus (Dim-

itrova *et al.*, 2009) contains more than one million words and combines texts from more than one domain – fiction texts and administrative texts (EU documents). Some parallel corpora in the OPUS collection (Tiedemann, 2009) include Bulgarian – medical documents by the European Medicines Agency, movie subtitles, and the SETimes news corpus.

Smaller and purpose-driven corpora, such as the *Nineteen Eighty-Four* (Multext-East) and the SEnAC and SEnFC corpora (SEE-ERA.NET), are tokenised, lemmatised, POS-tagged and aligned at sentence level. Annotation of larger corpora is usually limited to tokenisation, sentence splitting, and alignment, e.g. the OPUS collection, with the tendency to be extended to other levels of annotation.

Due to the limited amount of translations between particular pairs of languages, the interest in comparable corpora has been growing in the last decades. Still, only a small number of comparable corpora involve Bulgarian. The Multext-East comparable corpora with subcorpora of Bulgarian, Czech, English, Estonian, Hungarian, Romanian, and Slovene, include fiction and newspaper data (Dimitrova *et al.*, 1998). The Bulgarian-Croatian Comparable Corpus (Bekavac *et al.*, 2004) contains newswire texts, 393,000 tokens for Bulgarian and 1.3 million for Croatian. The Bulgarian-Polish-Lithuanian Comparable Corpus comprises fiction and electronic media documents balanced in size across the three languages (Dimitrova *et al.*, 2009).

This overview suggests that the existing Bulgarian corpora share most of the merits of the corpora compiled for other languages, and suffer from similar shortcomings, further aggravated by their smaller size and limited diversity, as well as the restricted availability of both monolingual and parallel data. On the positive side, most of the manually annotated corpora conform to the best annotation practices and have been employed in the development of various NLP applications for Bulgarian.

## 3 KEY FEATURES OF CORPORA

Apart from their use in traditional corpus linguistics and computational lexicography, contemporary corpora have been increasingly employed in developing language models, translation models and in training machine learning algorithms. However, despite the rapid de-

velopment of technologies and the vast amount of electronic data, the available corpora largely adhere to long-established tradition: they aspire to represent a balanced sample of language and for that reason constitute collections of carefully selected, often fixed-size, text excerpts. They are being explored with outdated methods and tools, which limits their use to extraction of concordances and collocations.

In the context of the dynamically evolving web and with more and more mono- and multilingual corpora becoming available, the traditional understanding of corpus design has been undergoing reconsideration. Some of the most prominent corpus features: size, balance, and representativeness (Xiao, 2010) will be discussed in the following subsections. We then proceed to propose a general approach for corpus development based on *automatic compiling*, *detailed metadata description*, and *multiple annotation*. This, we believe, will result in dynamic enlargement and efficient management of corpus data.

### 3.1 *Corpus size revisited*

A corpus large enough for empirical studies on language might not contain sufficient occurrences of specific and rare language phenomena for drawing statistically valid conclusions (Banko and Brill, 2001; Keller and Lapata, 2003; Kilgarriff and Grefenstette, 2003). Consequently, for building probabilistic models, larger amounts of data are needed, as large data, even if they are noisy, yield more reliable models than estimates based on smaller, limited datasets. After exploring the performance of a number of machine-learning algorithms for disambiguation when the size of the training corpus was increased from a million to a billion words, Banko and Brill (2001) concluded that the performance of any algorithm improves with data size, although the optimal data size varies with different algorithms[33]. This assertion is reflected in the rationale behind the web-as-a-corpus framework (Kilgarriff and Grefenstette, 2003), where the case is made for the necessity of making vast amounts of Internet texts available for processing and querying. Scientists have long since realised that the largest corpus is the web and that what primarily keeps Internet data

[33] An important insight made by Curran and Osborne (2002) in criticising Banko and Brill (2001) is that the benefits of large amounts of data are better experienced when size is combined with sophisticated statistical language models.

from becoming a real corpus is the lack of linguistically focused metadata and annotation.

The major size-related concern for corpus linguistics is how to define optimality – in terms of corpus size and in terms of sample size. The criterion for optimal corpus size aims at ensuring adequate coverage of lexical diversity, estimated with respect to wordstock, thematic domains, genres or language phenomena, while the criterion for text sample size takes into account the balance between texts, as well as their diversity. Several attempts at approximating an "ideal" size and structure, defined intuitively or empirically, have been made. Yang *et al.* (2000) try to estimate a corpus size that would be sufficient for obtaining the core vocabulary, while Chevelu *et al.* (2007) propose an algorithm for calculating an optimal corpus design that ensures coverage of a preset description of phonological attributes. What qualifies as optimal corpus size is still an open question, contingent on the particular linguistic or lexicographic task.

We shall illustrate the relation between corpus size and lexical diversity by a comparison between the "Brown" corpus of Bulgarian and FullBrown. The former consists of around one million words in 500 fixed-size samples of approximately 2000 words with adjustment to sentence boundaries; the latter includes the full-length originals of the BCB excerpts and totals 4.5 million words. The Bulgarian "Brown" corpus has 112,130 unique tokens, of which 61,162 (6.12% of all corpus tokens) appear only once. FullBrown contains 256,413 unique tokens and 130,230 tokens (2.89% of all corpus tokens) have a frequency of 1.

The early corpus tradition used a fixed size of the text samples to ensure balance and diversity of data and to avoid the skewing that might result from including large texts. Although limiting the size of samples is still appropriate for balanced and domain- and purpose-specific corpora, the above example shows that the inclusion of full texts contributes to language diversity and helps overcoming data sparsity.

The approach we adopt is based on two assumptions: that larger corpora are better suited to language analysis, irrespective of the particular task; and, that these resources, if properly documented and annotated, may also serve as a reliable source from which smaller, uniformly processed, different-sized subcorpora can be extracted, thus

eliminating the need for ad-hoc building of standalone fixed-structure corpora. Therefore we include the full versions of the texts in the corpus, as this allows us to extract comprehensive statistical metadata for the number of tokens, words, lemmas, clauses, sentences, and specific grammatical constructions that would enable further extraction of subcorpora, such as the one compiled for the development of the Bulgarian Sense-Annotated Corpus (BulSemCor; cf. Koeva *et al.* 2011).

### 3.2 *Balance and representativeness reconsidered*

The size of contemporary corpora comes at the expense of their structure, as they are usually created by collecting vast amounts of data at a fast rate. The predefined design criteria that used to be the organising principle of post-Brown corpora turn out to be empirically refuted and in need to be redefined in terms of the availability of various text types.

Representativeness is associated with the adequate coverage of the varieties of language use, while balance concerns the linguistically relevant distribution of texts across categories (Sinclair, 2005). Although these corpus features have been the focus of extensive study (Leech, 1991; Atkins, 1992; Biber, 1993; Sinclair, 2005; McEnery *et al.*, 2006), definitive qualitative or quantitative criteria for ensuring or evaluating them have not been convincingly established.

As a consequence, in traditional accounts, where the notions of balance and representativeness are defined in terms of supposedly relevant linguistic coverage and proportions of total language production, they remain tentative notions (Manning and Schutze, 1999; Kilgarriff and Grefenstette, 2003; Kupietz *et al.*, 2010). Moreover, defining them in this fashion is not adequate when it comes to the requirements posed to corpora by NLP. The new demands have called for a shift from compiling corpora that are carefully proportioned in terms of sample size and text types to expanding the quantity of the data.

Below, we attempt to redefine the relationship between size, balance, and representativeness in a data-driven perspective.

Representativeness is recast in terms of the range and diversity of text categories accompanied by enrichment of the sampling methodology. Since balance is hard to maintain for dynamic (constantly evolving) corpora, we suggest that instead of trying to maintain it for the

whole corpus, we extract different balanced subcorpora based on a large set of criteria (both preset or user-defined) such as time period, thematic domain, genre, author, density or distribution of certain language phenomena, etc.

We focus on amassing large amounts of texts that cover a variety of languages, media type, styles, domains, genres, and topics. The dynamic enlargement of the corpus, including the growth rate, the range of samples, and their quantity, is determined by the availability of texts on the web rather than by a preset model. Corpus structure is ensured through detailed metadata organised in a comprehensive classification of categories. The detailed metadata description allows for easy compilation of general, domain- and purpose-specific subcorpora with a fixed structure or predefined features. The metadata classification scheme is flexible, in order to match the new texts that are constantly being included in the corpus.

### 3.3 *Extended metadata and linguistic annotation*

Metadata describe the properties of the text samples in the corpus and are external to the text itself. Burnard (2005) emphasises the importance of metadata and the need for them to be as detailed as possible so that one may be able to determine the relevance of a given linguistic resource to one's own purposes. In the proposed framework, the classification suggested by Burnard (2005) is adopted as a baseline description of the text metadata and the annotation of the texts.

1. Editorial – information about texts in relation to their original source (source, author, date of publishing, etc. Here we include information about language, direction of translation, name of the translator, etc.);
2. Descriptive – classificatory information such as style, domain, and genre;
3. Administrative – documentary information about the texts and the corpus, such as its availability, revision status, etc.;
4. Analytical – various levels of annotation;
5. Statistical – number of tokens, words, general words, domain-specific words, lemmas, noun phrases, phrases, clauses, sentences, etc. In addition to Burnard's classification we include various statistical information.

Extralinguistic metadata about the texts (types 1-3) are built through a combination of automatic and manual techniques with increasing application of the former. Extralinguistic metadata are derived automatically from the HTML markup of the original files or with various heuristics based, for example, on domain-specific or genre-descriptive keywords. Statistical data (type 5) are calculated before or after annotation.

We represent the metadata scheme as a graph (Figure 1) where the nodes are associated with metadata categories and the arcs with binary relations between the nodes, such as *style*, *domain*, and *genre*, etc. For some metadata relations, for instance *style*, the metadata categories are predefined; for others, such as *author*, the categories are an open set. The representation is simplified, e.g. authorship of the text is recorded only once for all kernel and satellite samples in different languages. As a further advantage, graph representation allows flexible extension with new relations and categories and shows where merging or splitting categories is permissible. For example, it is possible to merge the metadata with a database of books' descriptions allowing us to automatically assign publishing dates or obtain translations of the title in different languages. Different "graph mining" algorithms – common subgraph, shortest paths, minimum spanning trees, connectedness, etc. can be used when extracting subcorpora of different types.

Linguistic annotation increases the value of a corpus by making it more *usable*, as various kinds of information may be extracted, more *multifunctional*, as the corpus may be used for different purposes, and more *explicit* with respect to the analysed information (McEnery *et al.*, 2006, p. 30). In our approach, we adopt and supplement the criteria set out by McEnery *et al.* (2006) as follows:

- **Multi-layered** – the more richly annotated the corpus, the broader its range of applications for research and applied studies. Corpus processing needs to cover and accumulate as many levels of linguistic annotation as possible.
- **Compliance with standards** in data formatting and representation of annotation. Unification of various tagsets and data formats, including encodings, is enabled through easy and reliable conversion.

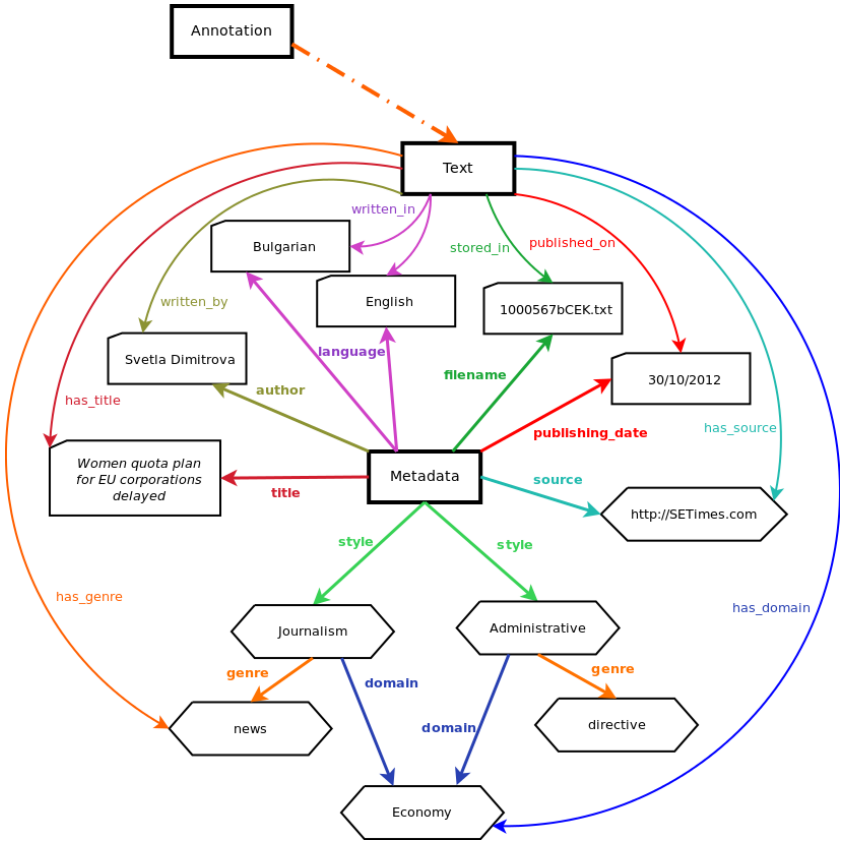



Figure 1: Example of the graph representation of corpus metadata.

- **Uniformity** – a common set of attributes and values for different languages and different media types – text, audio, image, video, and common techniques to manage (accumulate, combine, split, etc.) them. This will facilitate comparative studies and the application of language-independent tools.
- **Consistency** – as annotation of large amounts of texts in most cases is carried out automatically, it is necessary to provide means for validation and evaluation.

The following annotation principles are observed in general, both for manual and automatic annotation (Koeva *et al.*, 2010): the input text remains unchanged; the annotation is performed at consecutive stages and is accumulated as multi-level annotation; the annotation data are represented as attribute-value pairs. Each annotation level

is independent, may be accessed separately and merged with other compatible annotation schemes.

For the abstract representation of annotation an attribute-value formalism is used, in which the attributes are different types of linguistic categories (i.e., word sense, syntactic category, grammatical gender) associated with a set of values, for example *shtastliv* (en: lucky) has the following attributes and values: word sense: *'having or bringing good fortune'*, syntactic category: Adjective, gender: masculine. Ambiguity is not accepted in annotation, so each attribute is assigned a single value. The set of attributes depends on the language features and the granularity of the annotation and is thus open. Binary relations may also be defined between attributes (i.e., common noun – concrete, animate – human), making graph representation possible.

Fully-automated annotation is faster and more consistent although its precision might be lower than manual annotation. Our approach employs primarily automatic annotation, at any level of monolingual and parallel linguistic representation, and to whatever extent possible. Subcorpora with a concentration of a particular grammatical feature (such as singularia tantum) or language construction (such as noun phrases with a prepositional complement) may be extracted on the basis of the annotation.

### 3.4 *A unified corpus approach*

The need for high-quality monolingual and multilingual corpora further necessitates adjustment of corpus design principles in order to ensure a uniform treatment of monolingual and multilingual corpus parts, with all texts being compiled, documented, processed and accessed within a common framework.

The main source for corpus compilation is and will be the *Internet*, through downloading of readily available text collections or by web crawling. Modern corpus development has to be based on *automatic (and semiautomatic) collection, documentation and annotation* of monolingual and multilingual corpora, while manual work is mainly reduced to defining metadata and annotation schemes, annotation tagsets and the development of gold-standard corpora for training and testing. The requirements for corpus development can be summarized as follows:

1. **Collection** – predominantly automatic compilation of full-text corpus samples by means of web crawling based on preliminary manual and automatic web mining; automatic cleaning of junk formatting, and elimination of duplicates;
2. **Documentation** – detailed metadata extracted from the mark-up of the Internet documents, from the raw data by means of document categorisation and information extraction, and from the annotation by means of statistical processing;
3. **Annotation** – largely automatic linguistic annotation covering different linguistic levels and conforming to uniformity with respect to different languages and different media types.

The corpus design requires a clear-cut structure based on an explicit description of sample categories and explicit mapping between parallel samples in different languages. On the other hand, the corpus structure has to be flexible enough to allow for reorganisation around different categories or languages. This is ensured by a *detailed and consistent metadata documentation* of corpus samples.

Another point of discussion has been whether corpus design should be based on linguistic or extralinguistic criteria (Atkins, 1992; Sinclair, 2005). We subscribe to the idea that text sampling should be based on external criteria derived from the text's communicative function (style, genre, domain, source, year of publication, etc.), rather than on internal criteria that reflect the features of the language of the text (Clear, 1992), since the former afford a more reliable classification, and also to a large extent predetermine the linguistic features of the texts.

The principles for corpus design we have adopted are reflected in the following requirements:

1. Task-independent design ensuring as many monolingual and multilingual data as possible, illustrating different media types with their styles, genres, and domains.
2. Extensibility of the corpus through the inclusion of newly emerging categories attested in language production.
3. Flexibility and robustness of the design in order to facilitate reconsideration and restructuring of classificatory information about the texts. Carefully designed mechanisms for reorganising should

ensure that already included texts are not misclassified after the changes.

4. Adoption of mechanisms for accommodating texts that belong to multiple categories while any additional information is also properly stored and remains accessible.
5. Easy access to the relevant documents, including simple and efficient extraction of information, as well as grouping and regrouping of texts into subcorpora.

This corpus design is proposed in order to maintain simultaneously monolingual and multilingual parallel corpora and allow them to be compiled, preprocessed, annotated, evaluated and accessed through common or compatible tools, compliant with metadata and annotation description schemes, as well as with common (or convertible) annotation tagsets. This approach ensures standardisation, reusability and automation at all stages of corpora development and usage.

Corpus development at the present time needs to take into account the fact that the main purpose of corpora is natural language processing, and should try to answer this field's growing needs of reliable, linguistically enriched multilingual resources. Fulfilling such functions, corpora can successfully serve both corpus linguistics and computational lexicography, as detailed metadata and annotation facilitate the compilation of various domain- and purpose-specific subcorpora.

## 4 THE BULGARIAN NATIONAL CORPUS

The Bulgarian National Corpus is designed according to the outlined approach. The corpus contains a large variety of texts of different size, media type (written and spoken), style, period (synchronic and diachronic), and languages (Koeva *et al.*, 2012b).

The BulNC started as a monolingual general corpus and has been enlarged constantly, with the latest effort focused on the collection and annotation of parallel data and resulting in the Bulgarian-X Language Parallel Corpus (Bul-X-Cor). The parallel corpora in the BulNC consist entirely of texts that have a Bulgarian counterpart (original or translation) and one or more foreign-language correspondences that can also

be either original or translated. Both the Bulgarian and the foreign versions can be translations from a third language. Bulgarian serves as a pivot language for the parallel corpora, but any X-language is treated equally with respect to text type diversity, preprocessing, metadata scheme, general annotation principles, different levels of annotation, corpus quality evaluation and modes of access for (computational) research and implementations. The corpus may be used for tasks involving any pair of languages available in it. Applying the same principles and methodology used for the Bulgarian part of the BulNC and the Bul-X-Cor ensures, among other things, efficiency in terms of storage, as duplication of files between different parallel corpora is avoided and texts are stored and processed only once, unlike other corpora, such as the corpora in the OPUS collection.

### 4.1 *Compilation of the BulNC*

Three basic approaches have been applied in the compilation of both the kernel and the satellites:

1. **Using readily available text collections**. The kernel of the Bulgarian National Corpus was first compiled on the basis of the Bulgarian Lexicographic Archive and the Text Archive of Written Bulgarian, which together account for 55.95% of the corpus. Later, two domain-specific corpora from the OPUS collection were included, namely the EMEA corpus (medical administrative texts) and the OpenSubtitles corpus (film subtitles) representing respectively 1.27% and 8.61% of the kernel of the BulNC (see Figure 3). A large amount of news data in the Bulgarian Lexicographic Archive and the Text Archive of Written Bulgarian were provided by the publishers of various Bulgarian newspapers.
   The corpora were either obtained in plain text format or converted to it. Metadata were extracted automatically wherever possible, documented and verified manually in some cases. Full annotation was performed from scratch, even for already annotated texts (OPUS texts are tokenised and sentence-al-igned) to ensure conformity with the adopted principles and annotation standards.
2. **Manual compilation** by browsing the Internet. While being the primary approach in the past, manual collection has now been applied in a limited number of cases for small numbers of large

documents whenever the development of a focused crawler was deemed inefficient. Most of the previously developed corpora within the kernel of the BulNC were compiled manually, such as the Bulgarian "Brown" corpus. Recently, manual compilation has also been used for collecting parallel fiction texts in multiple languages, accounting for 3.70% of the kernel corpus.

3. **Automatic compilation** by web crawling is in general preferred. Some well-known and widely used approaches for automatic collection of corpora are adopted, tailored further to our specific needs and optimised with respect to the efficiency and precision of the output. Currently, automatically obtained subcorpora within the BulNC include a large amount of administrative texts, news from monolingual and multilingual sources, scientific texts and popular science (e.g., Wikipedia articles), altogether amounting to 30.47% of the Bulgarian kernel of BulNC.
   Manual and automatic web mining prior to the crawling process ensures crawling efficiency, as well as high-quality results when it comes to the validity of collected documents and the correspondence between parallel texts. As parallel resources involving Bulgarian are limited on the web, crawling was supported by direct targeting, automatic or manual, of the appropriate resources. The structure of source webpages is also considered when crawling, by applying either links traversal algorithms or URL templates as appropriate for each source.

Several crawling algorithms were examined (Paramita *et al.*, 2011) and the main technique chosen to be applied in the general crawler was the Breadth-First algorithm (Pinkerton, 1994). First, a generalised crawler with the main functionalities was developed. The crawler starts at the initial webpage of the respective collection of documents and either harvests the links recursively until the relevant pages containing the documents are reached, or uses URL templates to access the pages directly. In most cases, the websites containing parallel texts are very large[34] and a general (non-focused) crawler needs to process a very large amount of links and documents in order to select the relevant ones. The general crawler is therefore transformed into a focused crawler by adapting it to the structure of the source site

[34] `http://eur-lex.europa.eu`

as derived by automatic or manual web mining. The focused crawler either implements the link harvesting technique directly, or uses a particular set of URL templates specific for a given website. Next, the focused crawler ensures the relevance of the extracted documents by selecting only those texts that have Bulgarian equivalents. Some corpora are static and require a single run of the crawler, while others are dynamic (e.g., news websites) and need weekly or monthly crawls.

Procedures to verify the validity of the documents collected through automatic crawling are implemented: deletion of empty files obtained from either invalid or missing URLs, text size checks, and verification of encoding. Furthermore, genuine correspondence of parallel documents is checked by comparing URLs, file sizes, dates, etc. To conclude, focused crawling with preceding web structure mining (which considerably reduces the number of visited links) ensures high quality of the results and improves efficiency.

### 4.2 *Size of the Bulgarian National Corpus*

The kernel of the BulNC, consisting of all Bulgarian texts in the corpus, currently amounts to 979.6 million tokens. Although efforts have been made at ensuring the relative balance of the texts in terms of media type, written texts prevail significantly (91.11%), with spoken data representing only 8.89% of the tokens and being limited in variety – parliamentary proceedings, lectures, and subtitles.

At present, the Bul-X-Cor features 33 parallel corpora, the so-called satellites, adding up to 972.3 million tokens. The kernel and the satellites total 1.95 billion tokens altogether. Each parallel subcorpus within the Bul-X-Cor mirrors the structure of the kernel. Languages are not equally represented: the largest corpus is the Bulgarian-English parallel corpus (280.8 and 283.1 million words for Bulgarian and for English respectively); four other corpora comprise between 100 and 200 million tokens per language, sixteen parallel corpora are in the range of 30 to 52 million tokens per language, another seven in the range of 1 to 10 million tokens, and the rest are below one million, with the smallest ones being the Chinese, the Japanese and the Icelandic corpora with less than 50,000 tokens per language (Figure 2).

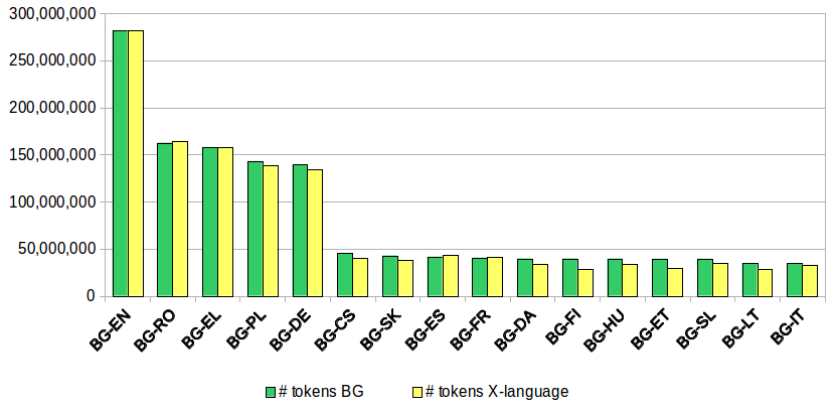



Figure 2: Largest Bulgarian-X language parallel corpora within the BulNC

### 4.3 *Structure of the Bulgarian National Corpus*

The structure of the corpus adheres to three main principles: explicit definition of categories, clear-cut structure and structure flexibility. The structure is not rigid in the sense that it is not predefined. The corpus samples are supplied with extensive metadata, facilitating the extraction of subcorpora with specific structure and features.

Language reflects communication in the following aspects: function and roles of the participants (style), thematic content (domain), and compositional structure (genre). The realisation of their interconnectivity is essential in building a good model for text description and classification. The design of the corpus is therefore based on the three basic classificatory features of style, domain, and genre.

Style is defined as a general complex text category, which combines the notions of register, mode, and discourse. The proposed approach does not rely on a particular linguistic theory of style, but is based on the analyses of Todorov (1984) and Halliday (1985), among others, who consider the intrinsic characteristics of texts in relation to external, sociolinguistic factors, such as the function of the communicative act.

Different terms are used in the existing literature: speech genre (Todorov, 1984), text type (Biber, 1989), register (Crystal, 1991). We have adopted the term *style* in the sense of Crystal (1969) with a more complex meaning that combines the notion of register (various degrees of formality of language (Trudgill, 1992)), media type (spoken or written) and discourse (function and characteristics of the com-

municative situation as reflected in the text). At present, the BulNC includes texts from six styles. Their distribution measured in number of tokens is presented in Figure 3.

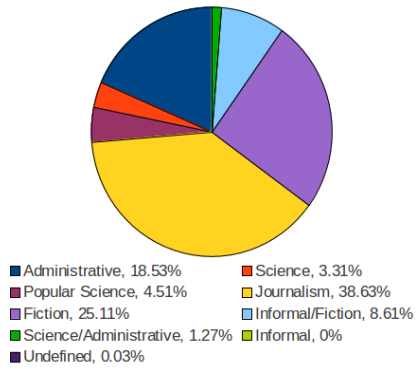



Figure 3:
Distribution of styles in the BulNC

A concise description of the text styles in the BulNC is presented in Table 1. Along with clear-cut styles, two complex styles are also included: Informal/Fiction and Science/Administrative. The former can be defined as informal texts within fiction (subtitles), and the latter as highly specialised (scientific, e.g., medicine) texts within the administrative style. Each of the complex styles exhibits features typical of both components and may share domains and genres with either of them.

Each style is subdivided into thematic domains. It is generally true that domains are style-dependent, although sometimes they are found across styles. For example, the scientific style is divided into categories according to scientific field, e.g., mathematics, economics, political science, etc. Some of the domains of journalistic texts are similar to those of scientific texts – politics, economy, etc.

The term genre also has multiple interpretations. For our purposes we accept the interpretation where genre is associated with the internal formal features of the text (Kress, 1993), although the notion is extended to all texts, both written and spoken. The classification of genres is also inconsistent across linguistic studies, and in particular in existing corpus descriptions (Lee, 2001). A general classification of genres based on style and a set of widely accepted genre types is used in the BulNC.

Table 1: Characteristics of styles in the BulNC

| **Style** | **Communicative situation** | **Function of the text** | **Features of the text** |
|---|---|---|---|
| Administrative | Between official bodies and individual or legal subjects; official, formal, indirect, written | Establishing, regulating and maintaining formal relationships | Relatively strict form and structure, repetitive, ambiguity is avoided |
| Science | Between researchers and other specialists; formal, indirect, written | Communicating scientific facts | Strict form and structure, extensive use of specialised (domain-specific) language |
| Popular Science | Between researchers and the wider public; not strictly formal | Communicating scientific facts in accessible and understandable form | Freer form and structure, less specialised language |
| Journalism | Mainly between journalists and the general public; indirect | Providing information, news and commentary | Relatively stable form and structure, some emphatic elements (e.g., in structure or lexis) |
| Fiction | Between authors and the general public; indirect | Entertainment and conveying aesthetic and moral values | Free and varied structure, consistent genre-specific elements |
| Informal | Personal communication; more often direct, informal | Conveying personal message, sharing information | Free and varied structure, diversity in linguistic expression |
| Informal/ Fiction (Subtitles) | Informal situations within fictional work | Same as fiction; within the fictional framework – personal communication | Characteristics of both styles |
| Science/ Administrative | Administrative situations within highly specialised scientific domains | Same as administrative | Characteristics of both styles |

| Style | Number of domains | Number of genres |
|---|---|---|
| Administrative | 11 | 16 |
| Science | 21 | 15 |
| Popular Science | 25 | 7 |
| Journalism | 19 | 12 |
| Fiction | 13 | 25 |
| Informal | (not represented) | (not represented) |
| Informal/Fiction (Subtitles) | 17 | 1 |
| Science/Administrative | 21 | 16 |

Table 2: Distribution of domains and genres across styles in the BulNC

Table 2 presents the number of domains and genres each style is divided into. Table 3 provides an example of the domains and genres for the Administrative style.

The distribution across domains of the samples in Bul-X-Cor is similar to the distribution in the kernel of the BulNC. The styles are represented as follows:

1. Administrative – EU legislation documents in 23 languages
2. Science/Administrative (Healthcare) – administrative documents from the European Medicines Agency in 23 languages
3. Journalism – news in 9 Balkan languages and English
4. Fiction – texts in Bulgarian, English, German, Romanian, Polish, Greek, Czech.
5. Informal/Fiction – subtitles of feature films, documentaries and cartoons in 29 languages.
6. Science – in Bulgarian and English.

Figure 4 illustrates the distribution of styles within the Bulgarian-English parallel corpus.

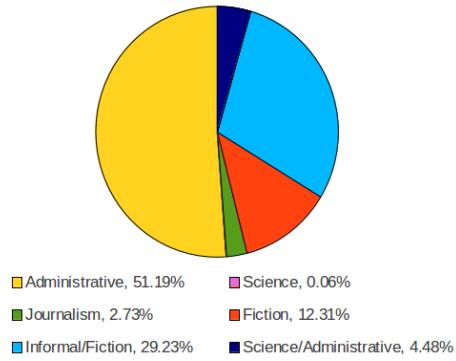
### 4.4 *Documentation and annotation*

The quality of corpus documentation and annotation has a major impact on the extent of its applications and usability. Therefore, great effort has been made to ensure that the documentation and annotation are accurate, well-structured and compliant with established stan-

Table 3:
Domains and genres for the Administrative style in the BulNC

| **Domains** | **Genres** |
|---|---|
| Politics | debates |
| Law | contract |
| Education | report |
| Economy | application form |
| Health | interview |
| Military | commentary |
| Culture and arts | correspondence |
| Sports | law |
| Ecology | plan |
| Social policy | programme |
| Relations | minutes |
| Undefined | certificate |
| | directive |
| | proceedings |
| | information |
| | document |
| | undefined |

Figure 4:
Distribution of styles in the Bulgarian-English parallel corpus



dards. Several levels of metadescription have been performed on the texts in the BulNC:

- Metadata documentation – metadata extraction and validation or metadata description of texts in any language;
- Monolingual annotation – processing of texts in any language at various linguistic levels;
- Multilingual annotation – alignment of parallel texts.

### 4.4.1 Text metadata

The metadata description of the texts in the BulNC is stored into 25 categories (Table 4) that are compliant with the established standards (Atkins, 1992; Burnard, 2005), although defined for the particular needs of the BulNC.

| | | |
|---|---|---|
| filename | path_to_file | date_added_to_corpus |
| author_info | author | translator_info |
| translator | text_info | title |
| year_of_creation | publishing_date | source_type |
| source | translated | medium |
| number_of_words | style | genre |
| genre_info | domain1 | domain2 |
| domain_info | notes | keywords |
| languages | | |

Table 4: Metadata in the BulNC

Metadata are mostly derived automatically, using two main techniques: (i) extracting information from the HTML or XML markup of the original files collected from the Internet, and (ii) keyword-based heuristics. HTML pages usually contain specifically tagged editorial information such as author, title, and date of publishing that are easily extractable from the HTML source.

Webpages within a website often contain similar texts, so focused crawling makes it easier to add classificatory information such as domain and genre. When classificatory information is not directly available, heuristics are applied to determine the domain and genre. One very simple example of using lists of domain-specific or genre-descriptive keywords is that if the title of a text contains a genre-specific word (e.g., *report*), it is assumed to denote the genre of the document.

The metadata are as detailed as possible in order to ensure easy text classification, corpus restructuring and evaluation, derivation of subcorpora based on a set of criteria (e.g., year of publication, domain). Some of the metadata categories, labelled with `_info`, are optional and contain additional details about the main category. A multiple domain description was also included to cater for the description of texts which have mixed domain features. So far, extensive metadata are provided for the Bulgarian and the English part of the BulNC, while the corresponding texts from the other languages share the common metadata (author, title, etc.) and inherit the classificatory information for style, domain, and genre.

#### 4.4.2 Monolingual annotation

Until recently, parts of the BulNC, such as the Bulgarian POS-Tagged Corpus, the Bulgarian Sense-Annotated Corpus, the Bulgarian-English Sentence- and Clause-Aligned Corpus were manually annotated (see Section 2.3), while lately we have focused on automatic annotation of larger portions of the corpus.

The linguistic annotation in the BulNC is divided into: (i) general monolingual annotation (tokenisation and sentence splitting), available for all languages, and (ii) detailed monolingual annotation, available only for the languages for which the respective tools and resources are available: Bulgarian and English. The detailed annotation so far includes morphosyntactic tagging (POS tagging and rich morphological annotation), and lemmatisation. The annotation of Bulgarian texts is further extended by including word senses, synonyms, hypernyms and `similar_to` adjectives, noun phrases, and named entities.

The Bulgarian texts are annotated using the Bulgarian language processing chain[35]. It integrates a number of tools (a regular expression-based sentence splitter and tokeniser, an SVM POS-tagger, a dictionary-based lemmatiser, a finite-state chunker, and a wordnet sense-annotation tool), designed to work together and to ensure interoperability, fast performance and high accuracy. The training of the Bulgarian tagger is based on the following parameters: two passes in both directions; a window of five tokens, the currently tagged word being

[35] `http://dcl.bas.bg/dclservices/`

in second position; 2- and 3-grams of words or morphosyntactic tags or ambiguity classes; lexical parameters such as prefixes, suffixes, sentence borders, and capital letters. Lemmatisation is based on linking the tagger output to the Grammatical dictionary (75 word classes to 1029 unique grammatical tags in the dictionary)[36], while a number of rules and preferences are applied to resolve the ambiguities. The finite-state chunker is a rule-based parser working with a manually-crafted grammar designed to recognise unambiguous phrases and to exclude pronouns, adverbs, and relative clauses as modifiers. The context-dependent rules provide annotation for phrase boundaries and heads.

Apache OpenNLP[37] with pre-trained models and Stanford CoreNLP[38] are used for the annotation of the English texts – sentence segmentation, tokenisation, and POS tagging. OpenNLP could be trained and applied for other languages as well. There are also some pre-trained models for a number of widely used languages (German and Spanish, among others). Lemmatisation of the English texts is performed using Stanford CoreNLP and RASP (Briscoe, 2006). As we aim at high quality and consistency of the annotation, we examine various systems for processing English and other languages.

Uniformity in annotation for Bulgarian and other languages is achieved in either of two ways: (i) annotation of raw data from scratch, applying equal standards and principles, or (ii) conversion of already existing annotation. In each case the tagset and conventions accepted for the BulNC are followed. The different tagsets are mapped to the Bulgarian tagset, but any language-specific annotation is preserved. The design of the Bulgarian tagset provides a uniform description of the inflexion of Bulgarian words and multiword expressions (Koeva, 2006) based on morphological and morphosyntactic criteria[39]. The tagset is mappable to the Multext-East morphosyntactic descriptions (Erjavec, 2004; Chiarcos and Erjavec, 2011), which are valuable as a unified framework for many European languages, although some disadvantages have been discovered with regard to the set of descriptions, both on a general and a language-specific level (Przepiórkowski and Woliński, 2003).

[36] `http://dcl.bas.bg/est/dict.php`
[37] `http://incubator.apache.org/opennlp/`
[38] `http://nlp.stanford.edu/software/corenlp.shtml`
[39] `http://dcl.bas.bg/en/BulgarianTagset_en.html`

### 4.4.3 Multilingual annotation

Multilingual annotation includes alignment at different linguistic levels, currently sentence and clause level. Alignment at sentence level is essential for all parallel resources and it is therefore required for all language pairs. High-quality sentence segmentation is an important prerequisite for the quality of parallel text alignment. The vast majority of the errors that occur in sentence alignment follow from inaccurate sentence segmentation. Two aligners have been applied for parts of the corpus: HunAlign (Varga *et al.*, 2005) and Maligna[40]. The alignment is based on the Gale-Church algorithm, which uses sentence-length distance measure and is largely language-independent. Other alignment methods, such as the Bilingual Sentence Aligner (Moore, 2002) and the use of bilingual dictionaries, are envisaged as well. The aligners take as input texts with segmented sentences and produce a sequence of parallel sentence pairs (bi-sentences). At present, alignment is performed and tested on the Bulgarian-English Parallel Corpus. A further step in parallel corpora processing is automatic clause alignment (Koeva *et al.*, 2012a), currently under way.

### 4.4.4 Annotation formats

Each raw text in the corpus is in plain text format. The annotation tools exchange data in the so-called vertical format, which is converted into an XML format and then stored in a MySQL database. In the vertical format, the tokens are separated by a newline and the annotation tags by a tab character. Each tool accumulates tags in fixed positions in one or several columns (for tags with a complex structure). Tags can be associated with a single token or a group of tokens.

```
new            TOK_LAT     new          A
technologies   TOK_LAT     technology   Np
.              TOK_FS<S>   .            U
```

The XML format also provides flexibility for representing various levels of annotation in both flat form (as sequences of elements) or hierarchical form (as nested elements, particularly useful for syntactic annotation). In the flat XML format adopted so far the text is represented as a sequence of `words` with associated attributes and their values that store the annotation information.

[40] `http://align.sourceforge.net/`

```
<word w="new" l="new" sen="11439" pos="A">
<word w="technologies" l="technology" sen="11439" pos="Np">
```

## 4.5 *General evaluation of the BulNC*

The monolingual and parallel parts of the BulNC can be evaluated from several perspectives, either quantitatively or qualitatively.

- **Quality of compilation methods**
  The quality of the crawling is ensured by implementing several techniques: manual and automatic data mining prior to crawling, development of a focused crawler for efficiency, as well as methods for verification of the results.
- **Statistical methods for qualitative analysis and evaluation**
  The strategy to gather a greater variety of word-grams and their distribution rather than to achieve balanced text category distribution is dominant. The aim is to employ statistical methods for qualitative analysis and evaluation, e.g., the proportion between the number of unique tokens / lemmas in the corpus and their frequencies / coverage / distribution within different (combinations of) styles, domains, and genres. It is assumed that variety in word distribution presupposes variety in text categories. For example, sparsity is evaluated through Zipf's law for frequency distribution and type-to-token ratios between old and new words (Goweder and De Roeck, 2001), and violation of Zipf's law may indicate data sparsity.
  Word sequence distribution (higher-order N-grams) may be combined with smoothing and skipping techniques (i.e., calculation of conditional probability based on different context) and with word similarity measures for automatic word clustering (Koeva *et al.*, 2012a). In that respect, the new data not only contribute by adding new lexical units, but also by supporting the saturation of the language model based on the previously collected lexical units.
- **Quality of metadata and annotation**
  Effort is made to ensure high-quality annotation in terms of accuracy, variety and coverage. On the average, in each metadata record in the BulNC 17.79 categories are non-empty (71.16%), a

figure that shows a good overall coverage for the description of the corpus texts.

The POS and grammatical tagger included in the Bulgarian language processing chain (Koeva and Genov, 2011) performs with a precision of 96.58%. The precision reported for the pre-trained model of OpenNLP used for the POS tagging of the English texts is 96.59%. Access to the processing tools is provided through a web service[41] or a web interface for asynchronous tasks[42].

### 4.6 *Access and applications of the BulNC*

The BulNC has been developed and expanded primarily to meet the needs of natural language processing. Still, the broad range of areas of application of the corpus makes it well-suited for public availability.

#### 4.6.1 Public access to the BulNC

As for the public access to the BulNC[43], we fully comply with Bulgarian and EU legislation concerning copyright and related rights. The law permits the use of copyrighted material for purposes of non-commercial scientific research and for education or private study. Where possible, we extract and store information about the source and the author's name and cite it accordingly. Several types of access to the corpus are provided: (i) download (limited); (ii) web search interface; (iii) collocation service; (iv) subcorpora selection; (v) frequency lists derived from the whole corpus or a given subcorpus.

Due to the inclusion of copyrighted material, the BulNC is not downloadable in full. For several style-specific subcorpora no redistribution limitations are in force, and these are available for download (registration required).

Like many of the large corpora presented in Section 2.1, the BulNC is supplied with a web interface for searching the corpus, as well as for building concordances and extracting collocations. The search system[44] (Figure 5) allows complex linguistic queries involving different levels of annotation combined in various ways. It is designed to support monolingual and parallel corpora in a uniform way. As compared

[41] http://dcl.bas.bg/dclservices/

[42] http://dcl.bas.bg/dclservices/admin/

[43] http://ibl.bas.bg/en/BGNC_access_en.htm

[44] http://search.dcl.bas.bg

to the CQL (Christ and Schulze, 1994), the implemented Designed query language (DQL) (Tinchev *et al.*, 2007) supports terms, such as: word – e.g. `word`; arbitrary word – e.g. `*{POS=A POS=ADV}`, relation – e.g. `word/F/`, and their combinations – e.g. `word/S/{POS=N}`. It is not restricted to a predetermined set of relations – at the moment queries for word forms, synonyms, hypernyms, and `similar_to` adjectives are supported. The atomic formulae allow both ordered and unordered queries, the latter being relevant for matching adjacent constituents with free word order, e.g., verbal clitics in Slavic languages. DQL is recursive and all Boolean combinations of formulae are formulae. This allows, among other things, disjunction of ordered queries, i.e., searching for paraphrases. The system also supports queries with regular expressions. For a given query the system retrieves matches in all documents regardless of language. Thanks to the alignment, the corresponding sentences in parallel documents are also accessible. The hits are paginated and the matches are highlighted. The user is able to view the detailed information for a given sentence in the hit set – the sentence metadata, its context, and correspondence(s) in the other languages.

The BulNC Collocation service employs the free NoSketchEngine[45], a system for corpora processing. The collocation service is a RESTful web service, supporting complex queries through HTTP and providing statistical information.

For instance, the query

```
http://dcl.bas.bg/collocations/?cmd=collocations&word=cat&cbgrfns=3td
```

returns statistical significance calculated with MI3, T-score, and logDice.

In accordance with our view that the corpus should allow for easy compilation of domain- and purpose-specific corpora compliant to a set of predefined criteria – e.g., synchronic, specialised, balanced subcorpora, we intend to provide a web interface for subcorpora selection, processing, and analysis. The extensive metadata ensures that a large set of criteria is available to cater for various research purposes and requirements. At present, we offer an offline request-based service for subcorpora selection and compilation of frequency lists[46].

[45] `http://nlp.fi.muni.cz/trac/noske`

[46] `http://ibl.bas.bg/en/BGNC_access_en.htm`

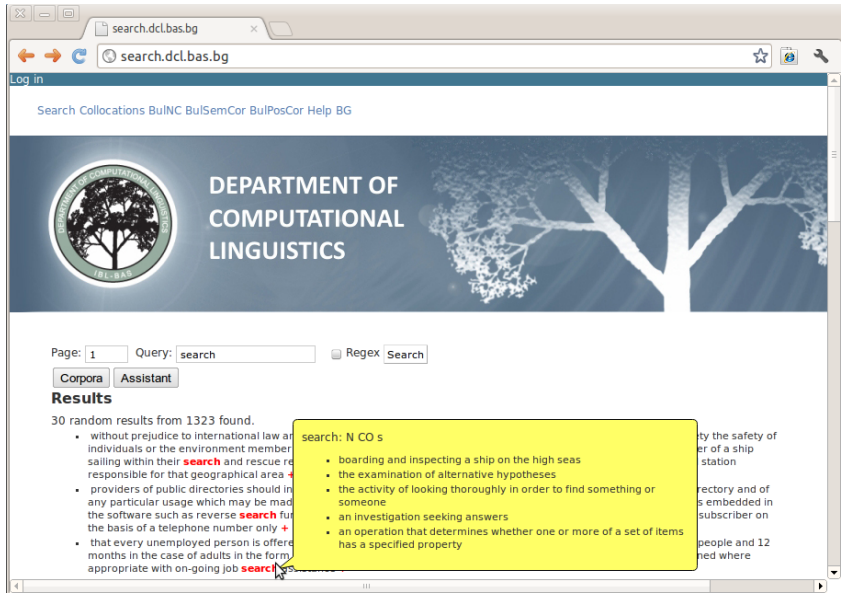



Figure 5: The BulNC search web service

### 4.6.2 Specialised subcorpora

Manually annotated subcorpora of the BulNC have been used as training and testing resources in numerous studies and NLP tasks, among them theoretical linguistic research, lexicological and lexicographic studies, POS tagging, semantic annotation and disambiguation, MWE recognition, parallel text alignment, clause segmentation and alignment, and many others.

For example, parts of the BulPosCor were used as training and test corpora in the creation of the SVM POS-tagger. The principal application of the BulSemCor is in the training and evaluation of a multi-component word sense disambiguation system. The corpus Wiki1000+, which contains Wikipedia articles (part of the *Popular science* style), includes 13.4 million words. Wiki1000+ was used for the purposes of recognition and classification of multiword expressions. The Bulgarian Sentence- and Clause-Aligned Corpus has been used for the purposes of parallel text alignment at sentence and clause level. It has served as a training resource in the development of a tool for clause alignment (Koeva *et al.*, 2012a). Several Moses[47] models (Koehn and Hoang, 2007) have been built on a large amount of par-

[47] `http://www.statmt.org/moses/`

allel data aligned at the sentence level in order to demonstrate the effect of syntactically enhanced parallel data (clause segmentation and alignment, reordering of clauses, etc.).

The applications of the BulNC and its subcorpora listed here are only a few examples of the numerous applications of the BulNC in the field of natural language processing.

## 5 CONCLUSION

In the context of the advance of technologies and the fast-growing amount of online information, the notions of text selection, balance, and representativeness can and should be reconsidered, shifting the focus from the theoretically grounded expectation for the distribution of text samples across different domains and genres to more sophisticated and flexible prediction based on calculations and estimations of language usage.

The paper has outlined the main concepts in corpus compilation with an emphasis on the key issues related to the use of corpora for the purposes of NLP research and applications. The attempt at redefining these concepts draws upon a discussion of the principles adopted in the compilation of large monolingual and parallel corpora for various languages. At the present time, large monolingual and multilingual corpora are constructed mostly by amassing text archives, repositories of documents, and bulks of texts available on the Internet.

Against this background, we propose a clear-cut approach for the compilation of a large multilingual corpus and demonstrate it in the context of the Bulgarian National Corpus. Our approach emphasises the extensive metadata and multi-level annotation of very large automatically collected monolingual and multilingual corpora, as well as the uniform treatment of multilingual content with respect to compilation, documentation, annotation, processing, and access.

## REFERENCES

Krasimira ALEKSOVA (2000), *Ezikat i semeystvoto. Kam metodikata za prouchvane na rechta v mikroobshtnostite (Language and the family. Towards a methodology for analysis of speech in micro social environments)*, Intervyu Press, Sofia.

Beryl Sue ATKINS (1992), Theoretical lexicography and its relation to dictionary-making, *Dictionaries: Journal of the Dictionary Society of North America*, 14:4–43.

Beryl Sue ATKINS, Jeremy CLEAR, and Nicholas OSTLER (1991), Corpus design criteria, `http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf`.

Michele BANKO and Eric BRILL (2001), Scaling to very very large corpora for natural language disambiguation, in *Proceedings of ACL 2001*, pp. 26–33.

Piotr BAŃSKI, Peter M. FISCHER, Elena FRICK, Erik KETZAN, Marc KUPIETZ, Carsten SCHNOBER, Oliver SCHONEFELD, and Andreas WITT (2012), The New IDS Corpus Analysis Platform: challenges and prospects, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2905–2911.

Piotr BAŃSKI and Adam PRZEPIÓRKOWSKI (2010), The TEI and the NCP: the model and its application, in *Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LRSLM2010)*, pp. 34–38.

Marko BARONI and Adam KILGARRIFF (2006), Large linguistically-processed web corpora for multiple languages, in *Proceedings of European ACL, Trento, Italy*, pp. 87–90.

Božo BEKAVAC, Petya OSENOVA, Kiril SIMOV, and Marko TADIĆ (2004), Making monolingual corpora comparable: a case study of Bulgarian and Croatian, in M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA, R. SILVA, C. PEREIRA, F. CARVALHO, M. LOPES, M. CATARINO, and S. BARROS, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, volume IV, pp. 1187–1190.

Douglas BIBER (1989), A typology of English texts, *Linguistics*, 27:3–43.

Douglas BIBER (1993), Representativeness in corpus design, *Literary and Linguistic Computing*, 8(4):243–258.

Ondřej BOJAR, Zdeněk ŽABOKRTSKÝ, Ondřej DUŠEK, Petra GALUŠČÁKOVÁ, Martin MAJLIŠ, David MAREČEK, Jiří MARŠÍK, Michal NOVÁK, Martin POPEL, and Aleš TAMCHYNA (2012), The joy of parallelism with CzEng 1.0, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Ted BRISCOE (2006), An introduction to tag sequence grammars and the RASP system parser, Technical report, University of Cambridge, Computer Laboratory Technical Report.

Lou BURNARD (2005), *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Metadata for corpus work, Oxford: Oxbow Books, `http://ota.ahds.ac.uk/documents/creating/dlc/index.htm`.

František Čermak and Vera Schmiedtová (2003), The Czech National Corpus Project and lexicography, in M. Murata, S. Yamada, and Y. Tono, editors, *Asialex '03 Tokyo Proceedings: Dictionaries and Language Learning: How Can Dictionaries Help Human and Machine Learning?*, pp. 74–80.

Atanas Chanev, Kiril Simov, Petya Osenova, and Svetoslav Marinov (2006), Dependency conversion and parsing of the BulTreeBank, in *Proceedings of the LREC Workshop Merging and Layering Linguistic Information, Genoa, Italy, 2006*, pp. 16–23.

Jonathan Chevelu, Nelly Barbot, Oliver Boeffard, and Arnaud Delhay (2007), Lagrangian relaxation for optimal corpus design, in *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pp. 211–216.

Christian Chiarcos and Tomaž Erjavec (2011), OWL/DL formalization of the MULTEXT-East morphosyntactic specifications, in *Proceedings of the Linguistic Annotation Workshop 2011*, pp. 11–20.

Oliver Christ and Bruno M. Schulze (1994), *The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual*, University of Stuttgart, Germany.

Jeremy Clear (1992), Corpus sampling, in G. Leitner, editor, *New directions in English language corpora*, Berlin: Mouton de Gruyter.

David Crystal (1969), *What is Linguistics?*, London: Edward Arnold.

David Crystal (1991), *A Dictionary of Linguistics and Phonetics*, Cambridge, MA: Basil Blackwell.

James R. Curran and Miles Osborne (2002), A very very large corpus doesn't always yield reliable estimates, in *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pp. 126–131.

Mark Davies (2010), The corpus of contemporary American English as the First Reliable Monitor Corpus of English, *Literary and Linguistic Computing*, 25(4):447–465.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimir Petkevic, and Dan Tufiş (1998), Multext-East: parallel and comparable corpora and lexicons for six Central and Eastern European languages, in C. Boitet and P. Whitelock, editors, *Proceedings of COLING-ACL'98: 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada*, pp. 315–319, San Francisco, Calif.: Morgan Kaufmann.

Ludmila Dimitrova, Violetta Koseska, Danuta Roszko, and Roman Roszko (2009), Bulgarian-Polish-Lithuanian Corpus – current development, in *Proceedings of the International Workshop Multilingual resources, technologies and evaluation for Central and Eastern European languages in conjunction with International Conference RANPL 2009, Borovec, Bulgaria, 17 September 2009*, pp. 1–8.

Tomaž Erjavec (2004), MULTEXT-East Version 3: multilingual morphosyntactic specifications, lexicons and corpora, in M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, and S. Barros, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, pp. 1535–1538.

Winthrop Nelson Francis and Henry Kučera (1964), Brown Corpus Manual, `http://icame.uib.no/brown/bcm.html`.

Roger Garside, Geoffrey Leech, and Tony McEnery (1997), *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman.

Voula Giouli, Nikos Glaros, Kiril Simov, and Petya Osenova (2009), A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts, in *Proceedings of Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, pp. 35–41.

Abduelbaset Goweder and Anne De Roeck (2001), Assessment of a significant Arabic corpus, in *Proceedings Workshop on Arabic Language Processing, 39th ACL, Toulouse*.

Atle Grønn and Irena Marijanovic (2010), Russian in contrast: Form, meaning and parallel corpora, *Oslo Studies in Language (OSLa)*, 2(1):1–24.

Michael Halliday (1985), *An Introduction to Functional Grammar*, Melbourne: Edward Arnold.

Frank Keller and Mirella Lapata (2003), Using the web to obtain frequencies for unseen bigrams, *Computational Linguistics*, 29(3):459–484.

Adam Kilgarriff and Gregory Grefenstette (2003), Introduction to the special Issue on Web as Corpus, *Computational Linguistics*, 29(3):333–347.

Adam Kilgarriff, Vojtěch Kovář, and Pavel Rychlý (2009), Tickbox Lexicography, in *eLexicography in the 21st century: new challenges, new applications*, pp. 411–418, Brussels : Presses universitaires de Louvain.

Jan Kocek, Marie Kopřivová, and Věra Schmiedtová (2000), The Czech National Corpus, in *Proceedings of EURALEX 2000*, pp. 127–132.

Philipp Koehn (2005), Europarl: A parallel corpus for statistical machine translation, in *Proceedings of MT Summit*, pp. 79–86.

Philipp Koehn and Hieu Hoang (2007), Factored Translation Models, in *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, June 2007*.

Svetla Koeva (2006), Inflection morphology of Bulgarian multiword expressions, in *Computer Applications in Slavic Studies*, pp. 201–216, Boyan Penev Publishing Center.

Svetla KOEVA (2010), Balgarskiyat semantichno anotiran korpus – teoretichni postanovki (Bulgarian semantically annotated corpus – theoretical concepts), in *Balgarskiyat semantichno anotiran korpus (Bulgarian semantically annotated corpus)*, IBL.

Svetla KOEVA, Diana BLAGOEVA, and Sia KOLKOVSKA (2010), Bulgarian National Corpus Project, in N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODJIK, S. PIPERIDIS, M. ROSNER, and D. TAPIAS, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pp. 3678–3684.

Svetla KOEVA and Angel GENOV (2011), Bulgarian language processing chain, in *Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg*.

Svetla KOEVA, Svetlozara LESEVA, Borislav Rizov RIZOV, Ekaterina TARPOMANOVA, Tsvetana DIMITROVA, Hristina KUKOVA, and Maria TODOROVA (2011), Design and development of the Bulgarian sense-annotated corpus, in *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politécnica de Valéncia*, pp. 143–150.

Svetla KOEVA, Svetlozara LESEVA, Ivelina STOYANOVA, Ekaterina TARPOMANOVA, and Maria TODOROVA (2006), Bulgarian tagged corpora, in *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, Sofia, Bulgaria*, pp. 78–86.

Svetla KOEVA, Borislav RIZOV, Ekaterina TARPOMANOVA, Tsvetana DIMITROVA, Rositsa DEKOVA, Ivelina STOYANOVA, Svetlozara LESEVA, Hristina KUKOVA, and Angel GENOV (2012a), Application of clause alignment for statistical machine translation, in *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), 12 July 2012, Jeju, Korea*.

Svetla KOEVA, Ivelina STOYANOVA, Rositsa DEKOVA, Borislav RIZOV, and Angel GENOV (2012b), Bulgarian X-language parallel corpus, in N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODJIK, and S. PIPERIDIS, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2480–2486, `http://www.lrec-conf.org/proceedings/lrec2012/pdf/587_Paper.pdf`.

Gunther KRESS (1993), Against arbitrariness, *Discourse and Society*, 4(2):169–191.

Marc KUPIETZ, Cyril BELICA, Holger KEIBEL, and Andreas WITT (2010), The German Reference Corpus DEREKO: A Primordial sample for linguistic research, in N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODJIK, S. PIPERIDIS, M. ROSNER, and D. TAPIAS, editors, *Proceedings of the*

*Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp. 1848–1854.

David LEE (2001), Genres, registers, text types, domains and style: clarifying the concepts and navigating a path through BNC jungle, *Language Learning & Technology*, 5(3):37–72.

Geoffrey LEECH (1991), The state of the art in corpus linguistics, in *English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik*, pp. 8–29, London: Longman.

Christopher MANNING and Hinrich SCHUTZE (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Tony MCENERY, Richard XIAO, and Yukio TONO (2006), *Corpus-Based Language Studies. An Advanced Resource Book*, Routledge.

Charles F. MEYER (2002), *English Corpus Linguistics. An Introduction*, Cambridge University Press.

Robert C. MOORE (2002), Fast and accurate sentence alignment of bilingual corpora, in *AMTA'02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pp. 135–144, London, UK: Springer-Verlag.

Graham NEUBIG (2011), The Kyoto Free Translation Task, `http://www.phontron.com/kftt`.

Cvetanka NIKOLOVA (1987), *Chestoten rechnik na balgarskata razgovorna rech (A Frequency Dictionary of Colloquial Bulgarian)*, Sofia: Nauka i izkustvo.

Monica PARAMITA, Ahmet AKER, Robert GAIZAUSKAS, Paul CLOUGH, Emma BARKER, Nikos MASTROPAVLOS, and Dan TUFIŞ (2011), Report on methods for collection of comparable corpora, ACCURAT – Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation.

Brian PINKERTON (1994), Finding what people want: Experiences with the WebCrawler, in *Proceedings of the First World Wide Web Conference, Geneva, Switzerland*, `http://thinkpink.com/bp/webcrawler/www94.html`.

Jan POMIKÁLEK, Miloš JAKUBÍČEK, and Pavel RYCHLÝ (2012), Building a 70 billion word corpus of English from ClueWeb, in N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODJIK, and S. PIPERIDIS, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 502–506.

Jan POMIKÁLEK, Pavel RYCHLY, and Adam KILGARRIFF (2009), Scaling to billion-plus word corpora, *Advances in Computational Linguistics. Special Issue of Research in Computing Science*, 41:3–14.

Adam PRZEPIÓRKOWSKI (2011), Linguistic annotation of the National Corpus of Polish, FDSL 9, `http://www.uni-goettingen.de/de/document/download/cbcf2e9ded91b3c41d0c460c31d1d9bb.pdf/nkjp.pdf`.

Adam Przepiórkowski, Marek Łaziński, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (2010), Recent developments in the National Corpus of Polish, in N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odjik, and S. Piperidis, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp. 994–997.

Adam Przepiórkowski and Marcin Woliński (2003), The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish, in *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC 2003), EACL 2003*, pp. 109–116.

Kiril Simov and Petya Osenova (2004), A hybrid strategy for regular grammar parsing, in M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, and S. Barros, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, pp. 431–434.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov (2002), Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Canary Islands, Spain*, pp. 1729–1736.

John Sinclair (2005), *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Corpus and text – basic principles, Oxford: Oxbow Books, `http://ahds.ac.uk/linguistic-corpora/`.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Daniel Varga (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 2142–2147.

Marko Tadić (2002), Building the Croatian National Corpus, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Canary Islands, Spain*, pp. 441–446.

Jörg Tiedemann (2009), News from OPUS – A collection of multilingual parallel corpora with tools and interfaces, in N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pp. 237–248, Amsterdam/Philadelphia: John Benjamins.

Tinko Tinchev, Svetla Koeva, Borislav Rizov, and Nikola Obreshkov (2007), System for advanced search in corpora, in *Literature and Writing in Internet*, pp. 92–111, Sofia: St. Kliment Ohridski University Press.

Tzvetan Todorov (1984), *Mikhail Bakhtin: The Dialogical Principle*, Minneapolis: University of Minnesota Press.

Peter TRUDGILL (1992), *Introducing Language and Society*, London: Penguin.

Dan TUFIŞ, Svetla KOEVA, Tomaž ERJAVEC, Maria GAVRILIDOU, and Cvetana KRSTEV (2009), ID10503 Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages, in *Scientific results of the SEE-ERA.NET Pilot Joint Call, Vienna*, pp. 37–48.

Dániel VARGA, László NÉMETH, P'eter HALÁCSY, András KORNAI, Viktor TRÓN, and Viktor NAGY (2005), Parallel corpora for medium density languages, in *Proceedings of the RANLP 2005*, pp. 590–596.

Ruprecht VON WALDENFELS (2006), Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment, in B. BREHMER, V. ZHDANOVA, and R. ZIMNY, editors, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pp. 123–138, München: Sagner.

Ruprecht VON WALDENFELS (2011), Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB, in Daniela M. and R. GARABÍK, editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20-21 October 2011*, pp. 156–162.

Richard XIAO (2010), Corpus creation, in *The Handbook of Natural Language Processing*, pp. 147–165.

Jia XU and Weiwei SUN (2011), Generating virtual parallel corpus: a compatibility centric method, in *Proceedings of the Machine Translation Summit XIII*.

Dan-Hee YANG, Pascual Cantos GOMEZ, and Mansuk SONG (2000), An Algorithm for Predicting the Relation between Lemmas and Corpus Size, *ETRI Journal*, 22(2):20–31.