ŁÓDŹ STUDIES IN LANGUAGE

Edited by
Barbara Lewandowska-Tomaszczyk

24

Stanisław Goźdź-Roszkowski (ed.)

Explorations across Languages and Corpora

PALC 2009

PETER LANG
Internationaler Verlag der Wissenschaften

Explorations across Languages and Corpora

ŁÓDŹ
STUDIES IN LANGUAGE

Edited by
Barbara Lewandowska-Tomaszczyk

Vol. 24

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Warszawa · Wien

Stanisław Goźdź-Roszkowski (ed.)

Explorations across Languages and Corpora

PALC 2009

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

ISBN 978-3-653-04563-5 (E-Book)
DOI 10.3726/978-3-653-04563-5

ISSN 1437-5281
ISBN 978-3-631-61677-2

www.peterlang.de

Table of Contents

Introduction

Ever since corpus linguistics entered the mainstream, it has become increasingly difficult to keep track of its most recent developments due to the sheer volume of corpus-based, corpus-driven or corpus-informed studies. For twelve years, the conferences known as PALC (Practical Applications in Language and Computers) and organized at the University of Łódź in Poland have served the international community of corpus and computational linguists by providing a useful forum for the exchange of views and ideas on how corpora and computational tools can be effectively employed to explore and advance our understanding of language. The conferences and the ensuing volumes have attempted to reflect the widening scope and perspectives on language and computers. The present volume is no different in that it documents new developments and explorations in these areas encompassing an array of topics and themes, ranging from national corpora and corpus tools through cognitive processes, discourse and ideology, academic discourse, translation, and lexicography to language teaching and learning. In keeping with the PALC tradition, it is our policy to publish contributions from both seasoned researchers as well as from colleagues who are recently initiated members of the corpus linguistics community.

The contributions are drawn from papers presented at the 7th *Practical Applications in Language and Computers PALC* conference held at the University of Łódź in 2009. The plenary speakers were Khurshid Ahmad (Trinity College, Dublin), Mark Davies (Brigham Young University), Ken Hyland (then at University of London), Terttu Nevalainen (University of Helsinki) and Margaret Rogers (University of Surrey).

This volume is divided into nine Parts, each Part being further subdivided into chapters.

Part One NATIONAL CORPORA provides overviews of three national corpora. First, Mark Davies (Brigham Young University) in his plenary paper *Semantically-Based, Learner-Oriented Queries with the 400+ Million Word Corpus of Contemporary American English* demonstrates the unique ways in which language learners can use the new Corpus of Contemporary American English to carry out and use semantically-based queries. František Čermák (Charles University, Prague) in *The Case of the Czech National Corpus: Its Design and History* introduces readers to the methodological issues of data acquisition and corpus design involved in creating the Czech National Corpus. The next two contributions are related to the National Corpus of Polish. Rafał L. Górski (Institute of Polish Language, Polish Academy of Sciences) in *The Design of the National Corpus of Polish* describes the proposed design of the National Corpus of Polish arguing that the corpus should reflect the perception of the language by the Polish linguistic community. Finally, Adam

Przepiórkowski (Polish Academy of Sciences) and Piotr Bański (University of Warsaw) in *XML Text Interchange Format in the National Corpus of Polish* describe and provide rationale for employing the XML encoding of texts within the National Corpus of Polish.

Part Two CORPUS TOOLS, INFORMATION AND TERMINOLOGY EXTRACTION comprises eight contributions. The first paper of this part is a contribution from Natalia Kotsyba (Warsaw University), Andrij Mykulyak (A. Sołtan Institute for Nuclear Studies, Warsaw), Igor V. Shevchenko (ULIF NANU, Kyiv) *UGTag: Morphological Analyzer and Tagger for Ukrainian Language* in which the authors describe the UGTag, a programme for morphological analysis and tagging of Ukrainian texts developed within the Polish-Ukrainian Parallel Corpus (PolUKR) project to support morphosyntactic annotation for the Ukrainian part of the corpus. The authors of the next seven papers which follow are Michal Křen (Charles University, Prague) and Martina Waclawičová (Charles University, Prague), who in their contribution *Database Framework for a Distributed Spoken Data Collection Project*, look at the main features of database system that is used in the Czech National Corpus for collecting recordings and transcriptions of authentic spoken Czech used in informal situations, Adam Przepiórkowski (Institute of Computer Science, Polish Academy of Sciences) and Grzegorz Murzynowski: *Manual Annotation of the National Corpus of Polish with Anotatornia* present the procedure of the manual annotation of a 1-million-word subcorpus of the National Corpus of Polish using a purpose-built tool, Anotatornia, Danuta Karwańska (University of Warsaw) and Adam Przepiórkowski in their paper *On the Evaluation of Two Polish Taggers* discuss the results of the comparison of two Polish taggers and the implications they carry for future taggers of Polish, especially the tagger, developed within the National Corpus of Polish, Tomas By (Centro de Linguística da Universidade Nova de Lisboa) in *Additional Comments on the Prolog Version of the Tiger Dependency Bank* updates information and provides more details on some of the methods used to verify that the word order disambiguation produces accurate results, Marcin Miłkowski (Polish Academy of Sciences) in his contribution *Automating Rule Generation for Grammar Checkers* decribes several approaches to automatic or semi-automatic development of symbolic rules for grammar checkers from the information contained in the corpora, Piotr Pęzik (University of Łódź) in *Providing Corpus Feedback for Translators with the PELCRA Search Engine for NKJP* introduces the PELCRA search engine for the National Corpus of Polish (PSEN) focusing on the usefulness of the tool in verifying the phraseology of translated texts, and finally Ewelina Kwiatek, Pius ten Hacken (Swansea University) in *Evaluating the Efficiency of MultiTerm Extract for Extraction of English and Polish Terms* investigate the efficiency of MultiTerm Extract, a component of SDL Trados 2007 to extract terms from English and Polish specialized corpora.

Part Three CORPUS-BASED LANGUAGE STUDIES includes the papers by Janusz Badio (University of Łódź), who in *What's in a "Duck"? A Corpus-based Study of Salience and Attention within Animal-Related, Denominal Verb,* attempts to find evidence for the claim that three processes of indexing, deriving affordances and meshing are used simultaneously in language understanding, Łukasz Grabowski (Opole University) in *Application of Parallel Corpora in Typological Investigations: the Case of Using English-Russian Parallel Subcorpus of the National Russian Corpus in Typological Study of Motion Verbs*, uses the English-Russian Parallel Subcorpus of the National Russian Corpus to explore the applications of this type of corpora in typological investigations, Milena Herbal-Jezierska (University of Warsaw) in *Corpus-based Morphology in the Czech Language* looks at various ways of using corpora in morpohological (especially inflexion) research in Czech; finally María Pérez Blanco (University of León) in her contribution *The Language of Evaluation in English and Spanish Editorials: A Corpus-based Study* discusses some of the linguistic features of newspaper editorials in English and Spanish.

Part Four COGNITIVE LINGUISTICS (Barbara Lewandowska-Tomaszczyk, University of Łódź and Paul Wilson, University of Łódź), *Culture Based Conceptions of Emotion in Polish and English* compares emotional dimensional structure of Polish and English native speakers and discusses the cultural bases to the findings.

Part Five groups papers which adopt A CORPUS-ASSISTED PERSPECTIVE ON DISCOURSE ANALYSIS AND IDEOLOGY. Cinzia Bevitori (University of Bologna at Forlì) in her paper *The Meanings of Responsibility in the British and American Press on Climate Change: A Corpus-Assisted Discourse Analysis Perspective* examines the ways in which the lemma 'responsibility' is used in a corpus of British and American press coverage of climate change in 2007. Mikołaj Deckert (University of Łódź): *Towards an Axiological Picture of the EU – Evidence from a Polish TV News Corpus* uses Polish broadcast news data to explore the axiological constructions of the personified European Union and its metonymic discursive forms. Katarzyna Fronczak (University of Łódź) in her contribution *Keywords and the Discourse of the Northern Ireland Peace Process, 1997–2007. A Case Study in the Election Manifestoes of the Democratic Unionist Party and Sinn Féin* adopts a keyword approach to carry out a textual analysis of language used in the election manifestoes focusing on the development and changes in the usage of words significant for the peace process. Joanna Kaim-Kerth (Jagiellonian University) in *Correspondence Analysis in Discourse Studies* employs multidimensional statistical analysis to demonstrate how such methods, particularly correspondence analysis, can be used in researching discourses focused on similar subject, i.e. values and authorities as well as visions of the family as perceived by different parliamentary fractions (understood as different

discourses). Finally, Inga Massalina (Kaliningrad State Technical University) in her study *Selected Cognitive and Discoursal Aspects of the LSP of the Navy* touches upon the cognitive and discourse aspects of the LSP of Navy.

ACADEMIC DISCOURSE is the topic of Part Six, which comprises three contributions. The first paper is a plenary contribution by Ken Hyland (University of Hongkong) *Corpora and EAP: Specificity in Disciplinary Discourses* in which the author draws on his own work conducted over several years into student and research genres to show how some familiar conventions of academic writing are employed by different fields. Silvia Cacchiani (University of Modena and Reggio Emilia) in her paper *Keywords and Key Lexical Bundles as Cues to Knowledge Construction in RAS in Economics* attempts to characterize disciplinary discourses by examining discourse signalling devices, and focussing on lexical bundles and keywords. In the last contribution of this Part *Conceptualising Spatial Relationships in Academic Discourse: A Corpus-Cognitive Account of Locative-Spatial and Abstract-Spatial Prepositions*, Christoph Haase (Chemnitz University of Technology) and Josef Schmied (Chemnitz University of Technology): investigate distributional properties of prepositions as heads of prepositional phrases that negotiate a mapping function between direct/literal and extended/metaphorical meaning when dealing with abstract concepts.

Part Seven deals with the theme of TRANSLATION and it includes eight contributions. It opens with a plenary contribution by Margaret Rogers (University of Surrey) *Translation Memory and Textuality: Some Implications*, who explores the use of computer-based tools in the translation of LSP (language for special purposes) texts and considers whether this has any implications for the nature of textuality. In the next paper, *Corpus Encoding and Integration for English-Spanish MT*, Alejandro Curado Fuentes (University of Extremadura) and Martin Garay Serrano (University of Extremadura) describe the compilation and encoding of the Spanish corpus and its integration on the database with the bilingual dictionary with a view to enhancing the Context-Based Machine Translation of English into Spanish. Dimitra Anastasiou (University of Limerick) in *Localisation, Centre for Next Generation Localisation and Standards* focuses on the description of a project called "Centre for Next Generation for Localisation" which currently runs in Ireland and she describes standards in terms of localization, especially the XLIFF standard. Hanne Eckhoff (University of Oslo), Dag Haug (University of Oslo) and Marek Majer (University of Oslo) in their contribution *Making the Most of the Data: Old Church Slavic and the PROIEL Corpus of Old Indo-European Bible Translations* report on ongoing work on the PROIEL corpus of old Indo-European New Testament texts, consisting of the New Testament in its Greek original and its earliest translations. Cécile Frérot (Université Stendhal Grenoble 3) in *Parallel Corpora for Translation Teaching and Translator Training*

Purposes explores the use of parallel corpora in designing a corpus-based translation course that is relevant from a translational standpoint and that is best suited to future professional translators, especially in terms of corpus-based translation tools. In the next paper, *The Combination of Comparable and Translation Corpora and How Translators May Benefit from it*, Marlén Izquierdo (University of Cantabria) continues the theme of parallel corpora by dealing with the combination of comparable and parallel corpora in the descriptive, functional analysis of languages in contrast and its positive impact for extending applications to translation. Marta Kajzer (Adam Mickiewicz University, Poznań) in *Translation of Eurojargon as a Source of Neologisms in Polish. A Corpus-based Study* presents a corpus-based analysis of Eurojargon translation as a potential source of neologisms in Polish. The last contribution in this section by Maria Tymczyńska (Adam Mickiewicz University, Poznań) *Community Interpreting in a Blended Environment – Student and Teacher Assessment* presents and discusses the application of Moodle, a free open-source online course management system, in the creation and implementation of practical interpreting courses in the Post-Graduate Programme in Community Interpreting offered at the Adam Mickiewicz University in Poznań, Poland.

Part Eight explores the theme of LEXICOGRAPHY. In the first contribution, Piotr Bański (University of Warsaw) and Beata Wójtowicz (University of Warsaw) in *New XML-encoded Swahili-Polish Dictionary: Micro- and Macrostructure* describe the structure of a new Swahili-Polish dictionary and some of the insights resulting from testing its electronic format. This is followed by Piotr Burmann's critical appraisal of a range of technical disctionaries in terms of their usefulness in the translator's work in *Insights into Selected Scientific and Technical Dictionaries Currently Available on the Polish Publishing Market.* Mirosława Podhajecka (University of Opole) in her contribution *Research in Historical Lexicography: Can Google Books Collection Complement Traditional Corpora?* demonstrates that the Google Books collection, a non-specialized textual resource can be applied successfully for research in historical lexicology and lexicography. The last paper in this section authored by Igor V. Shevczenko (ULIF NANU, Kyiv), Natalia Kotsyba (University of Warsaw), Kiryl Kurshuk (Hrodna University) *Towards the Creation of a Belarusian Grammatical Dictionary*, describes the process of creating the Belarusian grammatical word-inflexion dictionary, the first tool of the kind for this language, on the basis of linguistic similarities with the existing Ukrainian grammatical dictionary.

The last section in the volume, Part Nine LANGUAGE TEACHING AND LEARNING contains four contributions. First, Alex Boulton (CRAPEL–ATILF/CNRS, Nancy-Université) in his paper *Data-Driven Learning: the Perpetual Enigma* traces the evolution of DDL through the work of Tim Johns from 1984 up until his death in 2009, as well as in DDL studies by other

researchers. Then, Carmen Dayrell (University of São Paulo (USP) in *Anticipatory 'IT' in English Abstracts: A Corpus-based Study of Non-native Student and Published Writing,* explores the use of anticipatory *it* patterns (such as *it is found that* and *it is necessary to*) in English abstracts written by Brazilian graduate students from the disciplines of physics, pharmaceutical sciences and computing as opposed to abstracts of published papers from the same disciplines. Stefano Federici (University of Cagliari) in his contribution *Automatic Question Generation for E-learning* describes the "E-generation in E-learning" project that aims to evaluate and improve automatic question generation methodologies by means of analogy-based techniques to implement an automatic question generation system that best suites the needs of today e-learning platforms such as Moodle. Finally, Przemysław Krakowian (University of Łódź) in his paper *WEBCEF – The Project, Deliverables and Status Quo* looks at some of the issues and research questions which arose during the completion of WebCEF, a Socrates Minerva project in which web-based environments can be used for assisting the teacher in the process of evaluating spoken performance of language learners.

Acknowledgements

This volume and the preceding PALC conference would not have been possible without the support and encouragement from professor Barbara Lewandowska-Tomaszczyk. Special thanks are also due to dr Anna Cichosz, dr Piotr Pęzik, dr Jacek Waliński, Mr Mikołaj Deckert and Mr Łukasz Dróżdż for running the special workshops and sessions and for the overall efficient organization of the entire event.

Finally, we wish to acknowledge the financial support of the COST office (and personally dr Thekla Wiebusch of the CNRS, Paris), which enabled us to organize a COST Action A31 workshop at PALC 2009.

Stanisław Goźdź-Roszkowski

Part One

NATIONAL CORPORA

Semantically-Based, Learner-Oriented Queries with the 400+ Million Word Corpus of Contemporary American English

Mark Davies

Abstract: The architecture and interface for the Corpus of Contemporary American English (COCA) allow learners of English to carry out a wide range of semantically-oriented queries, including: 1) quick and easy collocates searches 2) comparison of collocates of two words (e.g. small/little) 3) comparison of collocates in different genres (e.g. collocates of "chair" in fiction and academic) 4) use of integrated thesaurus (entries for 60,000+ words) to see frequency of all synonyms (including by genre) and to create more powerful queries (e.g. all synonyms of "beautiful" + a synonym of "woman") and 5) customized wordlists (including hundreds or thousands of words in a semantic domain).

Keywords: English corpus, semantic acquisition.

1. Introduction

One of fundamental problems facing language learners is of course to acquire the semantic and pragmatic knowledge shared by native speakers of the target language. This involves such things as knowing:

- what words co-occur with a given word or phrase, which of course relates to native speakers' knowledge of what the word means and how it is used (i.e. "you can tell a lot about a word by the other words that it hangs out with")
- how the meaning and use of a word differs between genres
- the difference between related words, in terms of meaning and use
- how all of the words in a particular semantic field are related, in terms of frequency and distribution in different genres
  (See Schmitt 2000, Nation 20001, and Gardner 2007).

Corpora architectures and interfaces differ widely in terms of how much attention they pay to providing tools to answer questions such as these. It seems that sometimes these architectures and interfaces are oriented much more towards the interests of computer scientists and computational linguists than they are towards language learners.

In this paper, we will focus on how language learners can use the new Corpus of Contemporary American English (COCA)[1] to carry out and use semantically-based queries such as those listed above. As we discuss COCA[2], we will compare it to the four other architectures for large corpora that are currently available online for language learners:

1. Corpus Query Processor (CQP), as exemplified by its implementation in Sketch Engine (www.sketchengine.co.uk) (hereafter Sketch Engine)
2. Corpus Query Processor (CQP), as exemplified by its implementation in BNCweb (bncweb.lancs.ac.uk) (hereafter BNCweb)
3. VISL / CorpusEye (corp.hum.sdu.dk) (hereafter VISL)
4. Phrases in English (pie.usna.edu) (hereafter PIE)

BNCweb and PIE have only one corpus available – the British National Corpus (BNC), while Sketch Engine and VISL have several corpora – although no large corpora from the United States.

As we will see, both Sketch Engine and BNCweb offer fairly rich semantically-oriented queries. However, the range of semantically-oriented queries that are available with the architecture used for COCA is unique, and it is the only architecture that allows language learners to answer all of the types of issues shown above.

2. The composition of the corpus

Before discussing how the COCA architecture and interface can address this wide range of semantically-oriented queries from a learners' perspective, we should first briefly discuss the composition of the corpus, since of course the corpus data is only as good as the textual corpus on which it is based. For example, if we create a corpus that is based on just web pages and/or newspapers (the easiest types of materials to collect), then we will get a very skewed view of a given language. Ideally, we would want equal samplings from a number of widely divergent genres and registers, from genres as informal as

1 www.americancorpus.org

2 In this paper we refer to the "COCA" architecture and interface, as though it were particular to that one corpus. In reality, this architecture and interface have also been applied to a number of other textual corpora, such as the BYU-BNC, the TIME Corpus of Historical American English, the Corpus del Español, and the Corpus do Português. In this paper, however, we will focus on just the Corpus of Contemporary American English, and all of the examples are taken from that one corpus. Those who are interested in the more technical aspects of the corpus architecture and interface might consult Davies (2005) and Davies (2009a) for descriptions of earlier versions, and Davies (2009b) for a technical discussion of the current version.

spoken to genres as formal as academic, with a number of genres in between (cf. Biber *et al*. 1998).

In the Corpus of Contemporary American English (COCA), the corpus is divided almost equally between spoken, fiction, popular magazines, newspapers, and academic journals (see Davies 2009b for a more complete overview of the textual corpus). This composition holds for the corpus overall, as well as for each year in the corpus. As of August 2009, there are more than 160,000 texts in the corpus, and they come from a variety of sources:

- Spoken: (83 million words) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, *Oprah*, etc.).
- Fiction: (79 million words) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.
- Popular Magazines: (84 million words) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc.). A few examples are *Time*, *Men's Health*, *Good Housekeeping*, *Cosmopolitan*, *Fortune*, *Christian Century*, *Sports Illustrated*, etc.
- Newspapers: (79 million words) Ten newspapers from across the US, including: *USA Today*, *New York Times*, *Atlanta Journal Constitution*, *San Francisco Chronicle*, etc. There is also a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
- Academic Journals: (79 million words) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year.

There is no corpus of any language that is this large and which allows for a genre distribution this diverse. The British National Corpus has very good genre distribution, but is less than one fourth the size of COCA. The Cobuild / Bank of English corpus is somewhat larger than COCA (about 520 million words), but it is heavily weighted towards easily-available newspapers and contains very little spoken, fiction, or academic. The Oxford English Corpus is even larger, but it is based mainly on material from websites. In order to provide the best semantic and pragmatic information, we should have texts from a wide range of genres.

For a corpus this size, this range of genres is uniquely available in the Corpus of Contemporary American English.

3. Collocates: the size effect

As we will see, BNCweb and Sketch Engine provide useful insight into word meaning via powerful collocate-based queries of the BNC, as does COCA. At 100 million words, the implementation of the BNC in these two architectures seems like it would be quite adequate in terms of size for just about any collocate-based query. However, while 100 million words was huge when it was released in the early 1990s, it is beginning to look increasingly small, especially for collocate-based queries with lower-frequency words. Let us briefly look at just a few examples.

The following compares the collocates sets for given words in COCA and the BNC. The bolded number in the [BNC] and [COCA] columns shows the number of collocates that have a frequency of at least five tokens, within the specified span of words (e.g. 2 left, 0 right). The number in parentheses shows the overall frequency of the node word in the two corpora (e.g. the noun lemma *click* occurs 445 times in the BNC and 3145 times in COCA).

Table 1. Number of collocates in COCA and the BNC

node word (PoS)	collocate PoS / span	BNC **# collocates** (node freq)	COCA **# collocates** (node freq)
click (n)	adj 2L / 0R	**5** (445) double, sharp, loud	**26** (3145) loud, audible, double, sharp
nibble (v)	noun 0L / 3R	**2** (244) ear, bait	**28** (1194) edges, grass, ear, lip
crumbled (adj)	noun 0L / 3R	**0** (27) ---	**25** (446) cheese, bacon, bread, cornbread
serenely (adv)	verb 4L / 4R	**0** (83) ---	**30** (308) smile, float, gaze, glide

The node words were not selected on the basis of those that looked particularly good in COCA as opposed to the BNC. We simply looked for words with an overall frequency at a particular range in COCA, and then searched to see how many different collocates had a frequency of at least five tokens (within the specified span of words) in COCA and the BNC. As one can see, the difference is quite striking. Even though the 400+ million word COCA only has a little more than four times the number of words at the 100 million word BNC, it provides much richer collocational data for these lower frequency words. For

example, a word like *serenely* occurs only about four times as frequently in COCA as the BNC (which is to be expected in a corpus about four times the size), yet when one looks at moderately frequent collocates (five tokens or more), the difference is much more striking. In a smaller corpus like the BNC, the raw frequency of a node word might appear to be quite robust, but when it comes to finding collocates, the data often is too meager. As a result, in order to gain insight into the meaning and use of these lower frequency words, it seems clear that we need more than the standard 100 million word corpora of past decades.

4. Collocates: basic queries

Turning now to query types, perhaps the most basic type of information that a corpus architecture ought to be able to provide – in terms of meaning and usage – is collocational information. As the well-known saying in corpus linguistics points out, "you can tell a lot about a word by the other words that it hangs out with".

The VISL and PIE architectures can produce collocates for specific sequences of words (e.g. two word strings like *fast* [noun]), but they cannot find, for example, all nouns "near" *fast*. Sketch Engine and BNCweb are the two architectures, in addition to COCA, that do allow quick and easy access to full collocational information. With all three architectures, it is possible to define the collocates span (e.g. 2 words left or 5 words to the right of the node word), and to limit the collocates to a particular part of speech. In all three cases, the architectures are also quite fast – 1–3 seconds for a moderately frequent word like *catch* in the BNC (~15,000 tokens).

With COCA, for example, to find the most frequent nouns within three words after the verb *catch*, users would enter *[catch].[v*]* into WORD(S) (all forms of *catch* as a verb), *[nn*]* into COLLOCATES, and set the span to [0]–[3] (0 words to the left of *catch*, but up to three words to the right). They would then see results like the following:

Table 2. Collocates display (noun collocates of the verb *break*)

COLLOCATE	TOTAL		SPOK	FIC	MAG	NEWS	ACAD
HEART	1718		482	736	235	219	46
LAW	1478		739	134	196	335	74
NEWS	1379		616	211	224	225	103
SILENCE	992		174	540	119	91	68
RECORD	975		199	35	218	512	11
RULES	957		181	184	227	225	140
GROUND	884		114	96	228	325	121
STORY	786		519	54	94	94	25
TIME	780		164	248	193	115	60
WAR	710		206	87	153	153	111

The chart shown here is somewhat more complicated than a typical collocates chart. We see the frequency for each collocate in each of the five genres (and users can also see the frequency in each five year block since the early 1990s), but it is possible to also just see the overall frequency. The differences in collocates between different genres is something that we will return to in Section 7. In addition, we see here the raw frequency in each genre (color-coded by frequency per million words), but it also possible to see the actual normalized frequency in the chart as well.

COCA, Sketch Engine, and BNCweb also allow users to sort the collocates by "relevance" using Mutual Information score (MI) or other statistical tests. In addition, with each of these architectures it is possible to limit the results to just those collocates with a certain frequency or above a certain Mutual Information score. For example, the following are the most frequent collocates of the verb lemma *break* from COCA where the collocate occurs in the span [4 left / 4 right] at least 20 times, and the results are ranked by Mutual Information score.

Table 3. Collocates display: sorted by Mutual Information score

COLLOCATE	TOTAL		ALL	%	MI
LOGJAM	74		178	41.57	8.04
DEADLOCK	122		464	26.29	7.38
MONOTONY	70		346	20.23	7.00
COLLARBONE	74		445	16.63	6.72
STRANGLEHOLD	42		280	15.00	6.57

TABOOS	52	545	9.54	5.92
IMPASSE	78	881	8.85	5.81
SCUFFLE	30	354	8.47	5.75
BURGLARS	34	419	8.11	5.69
STALEMATE	61	806	7.57	5.58
LEVEES	64	861	7.43	5.56
BARRIER	394	5543	7.11	5.49

[TOTAL] shows the number of times that the collocate appears within the indicated span, [ALL] is the total number of tokens for that word in the corpus (e.g. 114 total occurrences of *red-handed* in the corpus, with or without *catch*), [%] is the percentage of tokens that occur in the span near *catch*, and [MI] is the Mutual Information score.

5. Collocates: more advanced queries

As we have seen, COCA, Sketch Engine, and BNCweb all allow for basic collocates functionality. What is the difference, then, between these architectures? The first difference is the range of node word and collocates pairs that the architectures allow. While Sketch Engine and BNCweb allow users to limit by part of speech and word form for the collocates, they are somewhat more limited than COCA, which allows all of the following:

Table 4. Types of collocate-based searches with COCA

NODE	COLLOCATES	SPAN (L/R)	EXPLANATION	SORT BY GROUP BY	EXAMPLES
LAUGH.[N*]	*	5/5	Any words within five words of the noun *laugh*	Percentage Collocates	hearty, scornful
[THICK]	[nn*]	0/4	A form of *thick* followed by a noun	Frequency Collocates	glasses, smoke
[LOOK] INTO	[nn*]	0/6	Nouns after a form of *look* + *into*	Frequency Collocates	eyes, future
[EYE]	clos*	5/5	Words starting with *clos** within five words of a	Frequency Both words	closed // eye closing // eyes

			form of *eye*		
[FEEL] LIKE	[*vvg*]	0/4	A form of *feel* followed by a gerund	Frequency Collocates	crying, taking
FIND	time	0/4	*Find* followed by *time*	Frequency Collocates	time
WORK/JOB	hard/tough/difficult	4/0	*Work* or *job* preceded by *hard* or *tough* or *difficult*	Frequency Both words	hard // work tough // job
[=PUBLISH]	[n*]	0/4	Nouns after a synonym of *publish*	Frequency Both words	publish // book issue // statement print // money
[=EXPENSIVE]	[[jones:clothes]]	0/5	Synonym of *expensive* followed by a form of a word in the *clothes* list created by *jones*	Frequency Both words	expensive / shoes, pricey // shirt
[=BOY]	[=happy]	5/5	Synonym of *happy* near a synonym of *boy*	Frequency Both words	happy // child, delighted // boy

With COCA, it is basically possible to look for "anything near anything else", including all synonyms of a given word, or any word in large "customized lists" that are created by the users via the web interface. The ability to look for "anything 'near' anything else" allows for very complex collocate-based queries, and the range of queries is unique to the COCA interface.

6. Collocates: word comparisons

COCA and Sketch Engine are the two architectures that allow for a powerful variation on regular collocates-based queries, one which involves comparison between words. Researchers have recognized the value of corpora in using collocates to tease apart slight differences between near-synonyms (e.g. *small* and *little*), or to provide insight into culturally-defined differences between two terms (e.g. *girls* and *boys*) (see, for example, Sinclair 1991 or Stubbs 1996). The architecture of COCA allows users to carry out searches such as this quickly and

easily, by comparing the collocates of two contrasting words or lemmas. For example, to compare the collocates of the adjectives *utter* and *sheer*, a user would simply select COMPARE WORDS, then enter *utter* in one search field and *sheer* in the other, and then select [nn*] as for CONTEXT. Finally, s/he might specify that the first word (*utter* or *sheer*) should occur at least 20 times with the given noun. The user would then see the following:

Table 5. Word comparisons (*utter / sheer* [NN*])

WORD 1 (W1): **UTTER** (0.30)

WORD	**W1**	**W2**	**W1/W2**	**SCORE**
DARKNESS	42	0	84.0	276.8
FAILURE	30	0	60.0	197.7
DESTRUCTION	26	0	52.0	171.3
DISREGARD	23	0	46.0	151.6
CONTEMPT	35	1	35.0	115.3
ABSENCE	15	0	30.0	98.8
AMAZEMENT	25	1	25.0	82.4
FOOL	12	0	24.0	79.1
DESOLATION	11	0	22.0	72.5
SILENCE	63	3	21.0	69.2

WORD 2 (W2): **SHEER** (3.29)

WORD	**W2**	**W1**	**W2/W1**	**SCORE**
NUMBER	226	0	452.0	137.2
VOLUME	187	0	374.0	113.5
FORCE	144	0	288.0	87.4
SIZE	241	1	241.0	73.1
WEIGHT	64	0	128.0	38.8
SCALE	62	0	124.0	37.6
LUCK	53	0	106.0	32.2
MAGNITUDE	51	0	102.0	31.0
AMOUNT	37	0	74.0	22.5
PLEASURE	73	1	73.0	22.2

This table shows that *sheer* occurs about 3.29 as common overall in the corpus as *utter*, and therefore all other things being example, it ought to occur about 3–4 times as frequently with any collocate than does *utter*. Conversely, *utter* should only occur about .30 times for each occurrence of *sheer*. In the case of *amazement*, however (#7 on the left side of the table), the collocate occurs 25 times as frequently with *utter* as with *sheer*, which is about 79 times as frequent as we would otherwise expect (taking into account its overall frequency, discussed above). In the case of *pleasure*, on the other hand (the last entry for *sheer*), it occurs about 22 times as frequently with *sheer* than we would otherwise expect. As one can readily see, the collocates for *utter* tend to be much more negative than those for *sheer*, and this points out an interesting semantic distinction that most non-native speakers of English would not otherwise be aware of (and perhaps not many native speakers either).

The following table provides a few additional examples of word comparisons that can be done with the corpus. [A] and [B] refer to the two words being compared, [Collocate PoS] shows the part of speech of the collocates, and the rightmost two columns show the collocates that occur with

either [A] or [B] much more than the overall frequency of either of these two words would suggest.

Table 6. Examples of word comparisons

[A]	[B]	Collocate PoS	Collocates with [A]	Collocates with [B]
[BOY]	[GIRL]	[j*]	growing, rude	sexy, working
DEMOCRATS	REPUBLICANS	[j*]	open-minded, fun	mean-spirited, greedy
CLINTON	BUSH	[v*]	confessed, groped, inhale	assure, deploying, stumbles
SMALL	LITTLE	[nn*]	amount, fee	while, luck
[ROB].[V*]	[STEAL].[V*]	[nn*]	bank, store	cars, money

In summary, the simple yet quick word comparisons that are possible with COCA would be of value to many different types of users (and similar word contrasts are possible with Sketch Engine as well). Linguists can quickly contrast synonyms and language learners can move beyond simple thesauruses to see more in-depth differences between words. And using COCA, even those interested in using corpora to investigate cultural studies, political science, and other social sciences (cf. Stubbs 1996) can quickly and easily compare how contrasting words (*Bush/Clinton*, *Democrats/Republicans*, *men/women*, *Christians/Muslims*) are used in contemporary American English.

7. Collocates: register differences

There is one type of collocates-based search that is possible only with COCA, and that is the comparison of collocates in two different sections of the corpus (such as genres or time periods) to see how word sense is a function of genre, or how a given word is changing in meaning over time. In this section, we will focus on genre-based differences.

Sketch Engine, BNCweb, and COCA all allow users to limit the query to one section of the corpus, such as Fiction or Newspapers-Sports. In COCA, however, it is possible to compare across two different sections. For example, the following table compares the collocates of *chair* in FICTION and ACADEMIC, and clearly shows the very different word senses in the two sections:

Table 7. Comparison of collocates by section

SEC 1: ACADEMIC

WORD	SEC1	SEC2	PM1	PM2	RATIO
DEAN	25	2	0.34	0.03	11.91
BOARD	76	8	1.04	0.11	9.05
COLLEGE	25	3	0.34	0.04	7.94
SECTION	39	5	0.53	0.07	7.43
COUNCIL	14	2	0.19	0.03	6.67
CONFERENCE	19	3	0.26	0.04	6.04
COMMITTEE	145	23	1.99	0.33	6.01

SEC 2: FICTION

WORD	SEC1	SEC2	PM1	PM2	RATIO
KITCHEN	197	2	2.83	0.03	103.35
LEATHER	209	3	3.00	0.04	73.10
LAWN	185	3	2.66	0.04	64.70
EYES	107	2	1.54	0.03	56.13
WINDOW	156	4	2.24	0.05	40.92
SWIVEL	137	4	1.97	0.05	35.94
ARMS	170	5	2.44	0.07	35.67

As can be seen, the collocates of *chair* that occur much more in ACADEMIC than FICTION are *dean*, *board*, *college*, etc., while those in FICTION but not ACADEMIC are *kitchen*, *leather*, *lawn*, etc. The tables show the frequency of each collocate with *chair* in the two sections (e.g. 197 tokens of *kitchen* near *chair* in fiction but only 3 tokens of *kitchen* near *chair* in academic). These then are converted to tokens per million words in the two sections (2.83 in FICTION, .03 in ACADEMIC), and the ratio figure (103.35) is the ratio of the normalized tokens per million figures for the two sections. As can be seen in this table, the data clearly show that in academic texts, *chair* refers to the position on a committee, whereas in fiction texts it refers to the piece of furniture.

With some modifications, the implementations of the BNC in BNCweb and Sketch Engine could conceivably allow for the same type of cross-genre comparisons, because of the way in which the BNC has been carefully constructed and annotated for genre and sub-genre. On the other hand, it would likely be very difficult to do this with "UK Web as Corpus" (ukWaC) corpus on Sketch Engine, because the architecture does not distinguish as clearly which web "genre" the texts belong to. Of all of the different corpus architectures, COCA is unique in the way in which it shows the relationship between genre and word sense.

8. Basic synonyms-based queries

To this point we have focused on collocates, which are of course one of the best ways of getting some sense of the meaning and use of words and phrases. However, there are two other powerful tools that are part of the COCA architecture and interface, and which are unique to this corpus.

The first feature relates to the integrated thesaurus in COCA. A standard printed thesaurus would show the following synonyms for *fast*: *quick*, *immediate*, *sharp*, *brief*, *sudden*, *rapid*, *swift*, *high-speed*, *abrupt*, *brisk*, *short-lived*, *speedy*, *fleeting*, *momentary*, *hasty*, *prompt*, and *hurried*. Obviously, however, some of these words are more frequent than others, and they would have a different distribution in different genres. Without this information, however, inexperienced language learners might end up sounding strange if they use *fleeting* or *momentary* much more than *fast* or *quick*. Language learners might also sound strange if they over-use a synonym in a genre where it is not appropriate, such as academic writing or in conversation.

COCA has an integrated thesaurus with entries for more than 60,000 synsets, which allows for powerful synonym-based queries. For example, users can enter a simple query like *[=fast].[j*]* (*fast* as an adjective), and then see the following (this is just a partial listing of all of the synonyms):

Table 8. Synonyms list (partial listing for the adjective *fast*)

	SYNONYM	TOTAL	SPOK	FIC	MAG	NEWS	ACAD
1	QUICK [S]	29005	6820	8057	6997	4993	2138
3	IMMEDIATE [S]	15606	2497	1445	3105	2948	5611
5	BRIEF [S]	15206	1845	3600	2790	2371	4600
7	FAST [S]	12713	1960	2524	4131	2736	1362
8	SUDDEN [S]	11345	978	5569	2125	1321	1352
10	RAPID [S]	10394	756	890	2377	1558	4813
14	SWIFT [S]	2817	393	907	714	510	293
15	HIGH-SPEED [S]	2671	269	124	1064	787	427
17	ABRUPT [S]	1886	115	610	408	278	475
18	BRISK [S]	1702	86	563	594	365	94
19	SHORT-LIVED [S]	1467	78	151	457	342	439
20	SPEEDY [S]	1388	184	196	493	353	162
21	FLEETING [S]	1330	93	472	366	191	208
22	MOMENTARY [S]	1068	51	504	185	92	236
24	HASTY [S]	940	77	346	182	160	175
27	PROMPT [S]	709	80	70	152	117	290

This table (which is about the most complex one that the user might see – most tables would be much more simple) contains a wealth of information. It shows all of the matching synonyms for the adjective *fast* in the thesaurus, along with their overall frequency and the frequency in each of the five main genres. Sketch Engine also has a "thesaurus-like" feature, but there are at least three important differences. The most basic difference is that the list of words are not really

synonyms per se, but rather words with shared collocates. For example, in Sketch Engine for the adjective *fast* it shows *slow*, *quiet*, *dangerous*, etc.

9. Comparison of synonyms across genres

A second difference between COCA and Sketch Engine is that COCA is the only corpus architecture that allows learners to see the frequency of all synonyms in the different genres, as are shown in the table above. With this information, users can see which synonyms are more formal or informal, and thus appropriate for different styles of speech. For example, the table above shows that *brisk*, *speedy*, and *hasty* are relatively less common in academic writing, but that *immediate*, *constant*, and *prompt* are relatively more common in that genre. Such information allows language learners to begin to develop some sense of which synonyms are most appropriate for a given target genre.

It is also possible to directly query to corpus to ask "which synonyms are more common in one genre than another?" For example, users could easily compare the synonyms of *smart* in newspapers vs. academic writing, by simply entering *[=smart]* for the word, and then selecting Newspapers for Section 1 and Academic for Section 2. They would then see that *ritzy*, *nifty*, *brainy*, *stylish*, *glitzy*, *chic*, and *trendy* are more common in newspapers, and that *intelligent*, *keen*, *clever*, and *shrewd* are more common in academic. Another example would be synonyms of *strong* in fiction and academic. In fiction, the synonyms are *beefy*, *burly*, *strapping*, *spicy*, *brawny*, and *pungent* (relating to people and foods), whereas in academic they are *effective*, *deep-seated*, *clear-cut*, *compelling*, *robust*, and *persuasive* – most of which refer to arguments.

10. Comparing collocates across a range of synonyms

A third and final difference with Sketch Engine, and one that adds real power to the synonyms feature in COCA, is the ability to include synonyms as part of more complex queries. For example, users can enter a query like *[=fast].[j*] [nn*]*, which would yield the following results. (These are just a handful of the more than 400 matching strings, and they are grouped by lemma (e.g. *fix* = *fix* and *fixes*), and include the frequency of that lemma string).

Table 9. Collocates of all synonyms of the adjective *fast*

SYNONYM	collocate	frequency
FAST	food	1178
QUICK	break	679
FAST	track	648

SYNONYM	collocate	frequency
IMMEDIATE	impact	224
RAPID	expansion	221
FAST	facts	205

QUICK	look	637
QUICK	question	529
QUICK	fix	511
BRIEF	moment	465
BRIEF	period	363
HIGH-SPEED	Internet	333
FAST	lane	275

IMMEDIATE	threat	195
SUDDEN	change	192
RAPID	pace	191
RAPID	succession	188
SHARP	decline	185
SHARP	increase	149

The power of a list like this is that users can quickly see which collocates occur with each of the synonyms. For example, language learner would see that native speakers talk and write about *brief moments*, *rapid succession*, *high-speed Internet*, *fast lane*, *sharp decline*, *immediate impact*, and *quick look*, but probably not *fast moments*, *brief succession*, *rapid Internet*, *sharp lane*, *quick impact*, or *rapid look*. Sketch Engine also allows users to click on words in the list of synonyms and to compare two words at a time, but COCA is the only corpus architecture that allows users to see all collocates with all synonyms at once. Such functionality allows users to move far beyond a typical thesaurus to compare competing words.

11. "Synonym chains"

Before leaving the topic of synonyms, we might mention one other very useful feature that is uniquely available via the COCA architecture and interface. As one can see in Table 8 above, each of the synonyms of a given word have an "[S]" after the word. Users can click on this to go from one synonym set to another, via a "synonym chain", and thus see an entire web of related words. For example, if users search for *beautiful*, they will see 18 synonyms, including *exquisite*. They can then click on the [S] after *exquisite* to see the synonyms of that word, including *delicate*, and from there to *sensitive*, and then to *mild*. And as before, for each of these synonym sets (as in Table 7), they can see a frequency-ranked listing of the synonyms, as well as see in which genre they are most common. All of this allows users to quickly and easily investigate a "web" of interrelated concepts and meanings, via a few clicks of the mouse.

12. Customized lists

One last feature of note related to semantically-based searches with the COCA architecture and interface is the ability to create one's own customized wordlists, and then seamlessly integrate these into the query syntax. There are two ways of

creating these lists. First, users can save a subset of the words or phrases from an existing search. For example, they could search for the synonyms of *beautiful*, or *crash*, or *money*, and then save just the synonyms that are of interest to them. Similarly, they can find the collocates of a given word, and then save some of these collocates in their own wordlist. They could simply create from scratch a wordlist, such as emotions (*sad*, *happy*, *worried*, *ecstatic*, etc.), colors (*blue*, *green*, *red*, etc.), or parts of clothing (*shirt*, *blouse*, *suspenders*, *hat*, etc.). In any of these cases, they simply create a name for the list and store it via the web interface under their chosen username.

These customized wordlists are saved in a database on the server, and can then be used a day, week, or year later as part of another query. For example, if a user *lingprof* creates a list for words related to emotions, s/he can then use these words as part of the query: *[r*] [lingprof:emotions] that*, to retrieve strings like *pretty worried that*, *quite sad that*, *extremely perturbed that*, etc. Likewise, these customized lists can be used as part of a collocates search. For example, the user *lingprof* might create a second list named *familyMember* (with *mother*, *mom*, *brother*, *uncle*, etc.), and then search for any *familyMember* within six words of one of the *emotions* words, e.g. *her aunt was quite happy to see that*, *when Dad is as angry as that*, *they were excited that Mom could be there*, etc. Again, the ability to incorporate user-defined lists as part of the query, as well as the basic corpus architecture, allows users to carry out quite complex semantically-oriented queries on the corpus. And again, this feature is not available with any other corpus architecture and interfaces.

13. Conclusion

As mentioned, one of the fundamental problems for language learners is the acquisition of the meaning and use of words and phrases. In order to do this efficiently, learners need to be able to quickly and easily find the collocates for a given word or phrase, see how the meaning and usage differs across registers, compare sets of collocates for two words to see differences in meaning between the words, compare collocates across a wide range of synonyms (hopefully all at one time), and see how the word compares in frequency and genre distribution with all related synonyms. Many corpus architectures – which are created by computational linguists or computer scientists – are oriented much more towards syntactic structure (parsing, complex regular expressions, etc.). Relatively few have semantically-oriented features like these, which are of real use to language learners. As we have seen, BNCweb does simple collocates quite well, Sketch Engine adds in a number of other features, but the COCA architecture and interface is perhaps the most advanced in terms of all of these different types of semantically-oriented queries.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., and E. Finnegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Davies, M. (2005). "The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation". *International Journal of Corpus Linguistics* 10: 301–28.

- (2009a). "Relational databases as a robust architecture for the analysis of word frequency". In: D. Archer (ed.). *What's in a Wordlist?: Investigating Word Frequency and Keyword Extraction*. London: Ashgate: 53–68.

- (2009b). "The 385+ Million Word Corpus of Contemporary American English (1990–2008+): Design, Architecture, and Linguistic Insights". *International Journal of Corpus Linguistics* 14: 159–190.

Gardner, D. (2007). "Validating the construct of Word in applied corpus-based vocabulary research: A critical survey". *Applied Linguistics* 28: 241–265.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

The Case of the Czech National Corpus: its Design and History

František Čermák

Abstract: A brief survey of needs and problems that led to establishment of a corpus institute and, subsequently, to the build-up of Czech National Corpus project are offered. These are followed by a presentation of a strategy adopted, data acquisition and their division, later on, into a number of corpora. Along the way, a number of problems had to be dealt with, out of which some attention is paid here to that of methodology, specifically such that would enable a representative shape of the contemporary corpora. Finally, a survey of existing corpora is presented and some open questions noted.

Keywords: Czech National Corpus, language data, methodology, corpus design, corpus representativeness.

1. General Remarks

Linguists have always suffered from data insufficiency, although they have only rarely admitted that this was the case. Reliable language data are a prerequisite and usual precondition for any information and subsequent conclusions that linguists are likely to draw, just as in any other science. Working with data has always been the mainstream in linguistics and Chomsky's stubborn and irrational contempt for any data hardly invalidates this general data necessity, a fact generally acknowledged. Despite of what he has mistakenly thought ("corpus data are skewed") it has always been quite clear that no one is able, for example, to write a dictionary from introspection only, i.e. the only approach he has subscribed to.

However, linguists may not have always been aware that they lack more data and reliable information being satisfied with what they had, a fact which is being gradually revealed only now, with new and better linguistic output based on and supported by better data. The old illustrious linguists like Otto Jespersen, a grand old man of English linguistics before the war, had been able to collect some 300 000 manual citation slips that he based all his grammars and books on. Today there is just no one willing to follow in his steps: having a corpus he/she does not have to. It used to be prohibitively expensive and time-consuming to collect large amounts of data manually, an experience familiar to anyone who has worked with citation slips from lexical and other archives trying to compile a dictionary or even, in the case of a student trying to write an essay required by his/her professor. To create such an archive of some 10–

15 million slips took many decades and many people (which was the case of Czech, too). Thus, there seemed to exist a natural quantitative limit that was difficult to reach and that was almost impossible to cross. Yet, the amount of information that could be acquired from such a limited archive was used for compilation of all dictionaries of the past, grammars and other reference books, including school textbooks we are still using today, all of them, seen from the contemporary perspective, sadly lacking in many respects, as they are neither contemporary, nor based on sufficient and convincing data. To sum up briefly: if you see one's language as a mirror of what is going on around us and compare the picture with old handbooks, then one did not have very good glasses looking at the language.

All of this has suddenly changed with the arrival of computers and modern very large corpora. For the first time in his/her personal life and in the history of the discipline as well, the linguist has multiples of previous amounts of data now, their flow being, in fact, so overwhelming and even staggering that he/she still has not got used to it and feels like a drowning person feeling a kind of embarassment as to how to handle the enorrmous amount of data one is facing. It is just beyond imagination, both for an old-timer as well as for a modern would-be lexicographer, to have to face, for example, some 83 000 occurrences of the word *člověk/lidé* (man/people) in Czech, a task neither of them has ever faced before. It is, to cool off some of the unwarranted enthousiasm, just not easy to properly use, almost a billion of words (as in Czech) since a lot of specialized work has to be done yet and many areas explored. This is just an illustration of how dramatically the data situation has changed for a linguist, while not all the consequences of this are fully grasped yet and solutions how to handle this found. It has become obvious now that the information to be found in this kind of data is both vastly better and representative of real usage than anything before and that the quality of information is proportionate to the amount of data amassed. Of course, this information has to be drawn from contexts, where it is usually coded in a variety of ways, implying that both the relevant texts and ways have to be found how to get at the information needed. Today, any concentrated work with a large corpus does cast a shadow over the quality and reliability of our present-day dictionaries and grammars making them problematic and dated. Obviously, there is a need for better resources and linguistic outputs based on these.

With modern corpora in existence, it is easy to see that these might be useful for many other professional and academic disciplines and quarters of life, including general public and schools, not only linguistics. After all, we are now living in the *Information Society*, as it has been termed, and it is obvious that there is a growing need for information everywhere. As there is, practically, no sector of life and human activity, no profession or pastime, where information is *not* communicated through and by the language, the conclusion seems

inevitable: the information needed is to be found in language corpora as the largest repositories of language. Should one fail in finding in corpora what he or she may need, then these corpora are either still too small, though they may have hundreds of millions of words already, or too one-sided and lacking in that particular type of language, as this happens to be the case of the spoken language. It is evident that there is no alternative to corpora as the supreme information source and that their usefulness will further grow. Corpora are an efficient shortcut and alternative to one's lifetime reading and listening the language and perceiving the information transmitted.

It may be also worth considering language in its proper perspective, as the first and most important attribute of a people and its culture: it seems that a corpus enabling to map the culture of its people might and should deserve a nation-wide attention and care of authorities. There is no better way how to spend one's money, in the long run, where culture and national heritage is concerned, since building a corpus amounts to constructing a permanent national monument.

Much of what has been been just said is general and holds for the Czech National Corpus project, too. Just like anywhere else, linguistic research in the Czech language had to be based on a data archive in past, catered for, in the old academic tradition, by the Academy of Sciences, namely through manually collecting language data on citation slips which have, over some eight decades, grown to reach some 12–15 million archive of excerpts. The decades-long excerption has been drastically cut down in the sixties when it was felt, for some reason, that enough data has been accumulated for a new dictionary of Czech, a decision difficult to understand from the contemporary point of view. Since then, most of the major linguistic work done has been based on this lexical archive including the compilation of a new large dictionary of contemporary literary Czech (*Slovník spisovného jazyka českého*) in four volumes (almost 200 thousand lemmas) which came out in 1960–1971. However, no extensive and systematic coverage of the language has been started ever since.

In the early nineties, a vague idea of a new dictionary of the Czech language has appeared in the hope that the new dictionary would capture the turmoil and social changes taking place after the Communist downfall, but the kind of data needed for this has been found to be non-existent. At the same time, it was becoming evident that the old manual citation slip tradition could not be resumed, especially one that would bridge a considerable data gap of over 30 years. My suggestion early in 1991–1992 was that a computer corpus be built from scratch at the Academy of Sciences to be used for such a new dictionary and for whatever it might be necessary to use it for, but the move was not exactly applauded by some influential people and old-timers at the Academy.

Yet, times have changed and it was no longer official state-run institutions but real people who felt they must decide this and also act upon their decision

and determination, no longer relying on problematic bureaucrats. Thus, thanks to the incentive of a group of people, a solution was found which took shape of a new Department of the Czech National Corpus which has been established in 1994 at Charles University in Prague (or rather its Faculty of Philosophy), introducing thus a base for the study subject of corpus linguistics as well (Čermák 1995, 1997, 1998). This solution was supported by a number of open-minded linguists who did feel this need, too. After the foundation of the Institute of the Czech National Corpus, all of these people continued to cooperate, subsequently as representatives of their respective institutions. Having joined forces, they now form an impressive cooperating body of people from five faculties of three universities and two institutes of the Academy of Sciences, altogether from ten institutions and more are still being addressed, especially in the task of oral data collection. Securing this kind of broad cooperation is now viewed as a lucky strike, indeed. Having gradually, though rather slowly at first, gained support, in various forms, from the State Grant Agency, Ministry of Education and from a private publisher, people have been found, trained and the Czech National Corpus project (CNC), being academic and non-commercial one, could have been launched. In the year 2000, the first 100-million word corpus, called SYN2000 has gone public and was offered for general use (Čermák 1997, 1998, Český národní korpus 2000), meeting with an enthousiastic welcome, mostly.

The general framework of the project is quite broad aiming at as broad a coverage of the Czech language as possible. Hence, more than one corpus is planned and subsequently built at the same time. Briefly, its aim is to cover the available bulk of the Czech language in as many forms as are accessible. The overall design of the Czech National Corpus consists of many parts, the first major division following the **III synchrony-diachrony** distinction where an orientation point in time is, roughly, the year 1990, for obvious reasons (the downfall of Communist régime and an enormous development and change of the language). Both these major branches are each split into the **(1) written**, **(2) spoken** and **(3) dialectal** types of corpora, though this partition, in the case of the spoken language, cannot be upheld for the diachronic corpora and there are problems with getting contemporary data from dialects, too. Yet, this is only the tip of the iceberg, so to speak, as this is preceded by much larger storage and preparatory forms our data take on first, namely by the **I Archive of CNC** and **II Bank of CNC**.

Let us now have a look at a brief outline that each language item (form) has to go through, listing main stages that the data go through before reaching their final stage and assuming the form that may be exploited. Of course, everything depends on the laborious zero stage of (**0**) **Text Acqusition** being finished, in which texts are gained from the providers mainly, which is not really easy and smooth as one would wish, often depending on the whims of individual

providers, legal act of securing their rights and physical transport of the data finally obtained. The Czech National Corpus Institute gets some texts either freely or on the basis of a modest fee, which has to be always supported by the consent of the original providers. There is no need stressing the fact that this is actually the easy way how to get the electronic texts. Two other ways, fortunately somewhat smaller in extent but much more laborious, expensive and labour-costly, are those of text-scanning (especially old texts, but also the authors the CNC did not have in its entirety) and of recording combined with manual transcription (the case of spoken corpora).

The first text format to be based on the data when they are acquired, in fact a variety of them, is stored in the **(I) Archive of CNC**. The Archive is constantly being enlarged and contains, at the moment, almost two billion words in various text forms. All of these texts are gradually converted, cleaned, unified and classified and, having been given all this treatment, they flow then into the **(II) Bank of CNC**. Thus, the Bank of CNC is a repository of raw but unified and "clean" texts prepared for any further treatment. A note has to be made about the conversion, however. This has to face the rich variety of formats publishers prefer to use and implies, in many cases, that a special conversion programme has to be developped allowing for this, though this is not always simple and reliable, such as with the popular and problematic pdf format. Of course, the cleaning of texts does not mean any correction of real texts, as these are sacrosanct and may not be altered in any way. Hence, efforts are made to find and extract **(1) duplicate texts** or large sections of them which, surprisingly and for a number of reasons, are found quite often.

Then, **(2) foreign language paragraphs** have to be identified and removed, these being due to large advertisements, articles published in the Slovak language, English, etc.

Finally, **(3) most of non-textual parts** of texts, such as numerical tables, long lists of figures or pictures are taken out, too (eg. stock-exchange columns of figures or sporting events tables).

So treated, each text gets, then, a modified **SGML (XML) format** containing an explicit and detailed information about the kind of the text, its origin, classification etc., including information about who of the staff of CNC is responsible for each particular stage of the process.

It is obvious that to be able to do this and achieve the final text stage and shape in the Bank of CNC, one has to have a master plan designed showing what types of texts should be collected and in what proportions. While more about this will be said later, it is necessary now to mention that this plan has been implemented and recorded in a special **(4) database** the records of which are mirrored in the corpus itself.

At this stage, after an elaborate weighting, selection, tagging and lemmatization (more about that later), some texts meeting the demands are

selected and proclaimed to be a **(III) corpus** that is given a name and, usually, made public on the web (www.korpus.cz). At the moment, there is a number of such corpora made available and more are being prepared (see part 6).

Originally, CNC was served by a comprehensive retrieval system called *gcqp*, which is based on the Stuttgart *cqp* programme. This has been considerably expanded (by P. Rychlý, a member of a partner team) and exchanged for a sophisticated graphic interface **Bonito** supported by Windows, although it is Linux-based, of course. Being now part of a client-server (**Manatee**), it offers a rich variety of search functions and facilities, including a possibilty to define one's own (virtual) subcorpus (see korpus.cz). This corpus manager is free and is being used by several universities and institutions.

2. Data and Resources

It is obvious that only the data that were available could be used first, the financial aspects of their acquisition being important, although most data are now given freely to the CNC on the basis of a prior contract with the provider (almost 300, of different nature, including publishers, newspapers, a number of private institutions, etc.); however, there are still some rare cases when the data have to be paid for.

Next to these electronically available texts, some texts that are not available, have to be either scanned into the computer (using OCR programmes, mostly FineReader) or manually recorded in the case of oral data. This means that a broad net of collaborators, mostly students, from virtually all major regional universities are secured and asked to record, for a small fee, their talk which should be as typical and spontaneous as possible involving a generalized cross-section of speakers.

3. Strategies Adopted

Realizing that not all familiar types of language are used in the same degree and that data for them are sometimes difficult to obtain, a decision to arrive at some kind of representativeness of most language types was adopted rather early. Being based on discussions and three subsequent stages of research in the domain of the written language (Čermák 1997, Čermák, Králík, Kučera 1997, Králík, Šulc 2005), the idea of representativeness has been oriented toward a general and broadly used vocabulary with the primary though not the only aim of eventually establishing a basis for a new general dictionary of Czech. With the stress laid only on language reception (i.e. the degree to what passive users have been exposed to the language, i.e. readers only) the reseach has offered a

balanced quantified picture, the only one that is available and based on research in any language, in fact. It is impossible here to specifically argue and substantiate every single item, which, after a further research, landed somewhere in a rich network that data have been classified into having now no less than a hundred categories in several strata. Thus, CNC might be now called to be a representative corpus which has been carefully planned from scratch. It is to be realized here that such a corpus becomes a referential and proportionate entity anyone can refer and come back to, a thing which Internet will never be. This also stands in sharp contrast to that type of corpora where any available text, preferably newspapers, is accepted, amassed into an amorphous entity and called a corpus. These rather spontaneous corpora do rely on a seemingly infinite supply of texts and the philosophy of great numbers, hoping somehow that even the most specific and specialized information might find its way into it eventually; a new version of this is to be seen in the blind and problematic reliance on the Internet. For many reasons, this could not be the Czech philosophy. Hence, these figures arrived at by this reaserch, which should be further scrutinized and revised, of course, were then being used for the fine-grained construction and implementation of the synchronic corpus SYN2000. The overall structure is this:

IMAGINATIVE TEXTS	**15%**
Literature	**15%**
Poetry	0,81%
Drama	0,21%
Fiction	11,02%
Other	0,36%
Transitional types	2,6%
INFORMATIVE TEXTS	**85%**
Journalism	***60%***
Technical and Specialized Texts	***25%***
Arts	3,48%
Social Sciences	3,67%
Law and Security	0,82%
Natural Sciences	3,37%
Technology	4,61%
Economics and Management	2,27%
Belief and Religion	0,74%
Life Style	5,55%
Administrative	0,49%

Let me make a single note at least, instead of many that could and should be discussed here. Any comparison with the few available data of this kind from elsewhere is problematic, as it obviously depends on how various subcategories are defined. To give an example, there is no consensus as to what might be viewed as Leisure, for example, the term used by the BNC and represented there by over 15%, which must overlap with what is termed Life Style in CNC (although its representation happens to be three times lower).

However, this general framework has recently been modified by the last research into libraries, data by publishers etc. so that the main division has been slightly shifted in favour of more literature, giving the definite tripartition as:

Literature	***40%***
Journalism	***33%***
Technical and Specialized Texts	***27%***

To give an idea how this is projected, in part, into the data annotation and the database, the following distinctions are used that every text has to fit into:

1 – type of corpus: synchronic, diachronic, spoken; parallel
2 – type of text: informative, imaginative or a mixture of both
3 – type of genre: different in specialized and non-specialized language (made of some 60 different types, e.g. drama, novels..., music, philosophy, industry, sport, ... religion, etc.)
4 – type of subgenre (such as text-book, criticism, encyclopedic, etc.)
5 – type of medium (such as book, newspaper, script, occasional, etc.)
6 – sex of the author if known (including a team)
7 – language (in the case of foreign language texts)
8 – original language (in the case of translation)
9 – year of publication
10 – name of the author if known or that of the translator
11 – name of the text/work
12 – identification of the part of a larger work/text

Somewhat later, it has been realized that the spoken language needs some master plan, too, which, obviously, must be quite different from that used for the written language. This has been based (Čermák 2007, 2008) on the idea of prototypicality of the spoken texts in the sense that protypical means different in all the aspects identified and not found in the written one. At the same time as it is, for example, really very easy to record TV or radio broadcasts that are often nothing more than written texts spoken for the microphone, it has been decided to go first after what is really different and prototypical, namely spontaneous dialogues, usually between friends. Admittedly, these are the most difficult to

get, too, so they seem to be a good start. The aspects suggested and used include:

PLUS (+)	**MINUS (-)**
a. Origin of the text:	
1 spoken (i.e. original)	– read
2 dialogue (i.e. original, typical)	– monologue
b. Interpersonal, sociological relationship of partners and physical situation:	
3 proximity of partners (friends, family)	– no proximity
4 equality of partners	– unequality
5 private (non-public)	– public
6 informal	– formal
7 interactive	– unidirectional
8 present	– distant (e.g. phone)
9 non-multiple (one-to-one)	– multiple (one-to-many)
c. Topic/situation approach:	
10 spontaneous (unscripted)	– prepared (more or less scripted, prepared)
11 casual (informal)	– regular/official
d. Awareness of the recording:	
12 not aware	– aware

In this framework, a fully and **prototypically spoken text** has all of the following plus parameters, such as a talk (conversation) between friends:

+*spoken(1)*,
+*dialogue(2)*,
+*proximity(3)*,
+*equality(4)*,
+*private(5)*,
+*informal(6)*,
+*interactive(7)*,
+*present(8)*,
+*non-multiple(9)*,
+*spontaneous(10)*,
+*casual(11)*,
+*not aware(12).*

Obviously, texts having all 12 plus parameters (PLUS features) may be viewed as the core, hence as a part of prototypical corpus itself.

4. Technical Aspects and Linguistic Treatment of Texts

Those **technical aspects** that are related to acquisition, conversion, cleaning, filtering and unification of data shaping them the way they can be further used have already been mentioned above (in 1). However important and time-consuming, they are performed mostly automatically now and are finished when raw data are ready for further use. However, they do require that a lot of specialized procedures has to take place before this, including tokenization, splitting texts into future units (i.e. very large texts), marking paragraphs in them, sentences and word tokens, within the SGML format. Finally, the SGML head is added in this series of technical steps. Should these texts be now used, one would only get pure textual type information offered by sequences of forms.

However, many linguists wanting more have introduced at least some **linguistic mark-up**, consisting mostly of **tagging** of word forms of texts and a subsequent **lemmatization**, both being automatic procedures that are dependent on a prior existence of a national reliable tagger and lemmatizer. For an inflected language like Czech this is easier said than done, however. Both procedures, based originally on a small training corpus that had been manually tagged, have been fighting ever since with never-ending problems, however. Due to a number of reasons, the limited size of the training corpus, enormous homonymy of forms in Czech, problems with mistaken statistical guessing of morphological values of text forms, etc., a great amount of work and effort is still going into this business requiring efforts of several people. It is now evident that, still being far from any real end (success rate being now around 95%), a combination of both methods will slowly improve the output in both procedures, namely the statistical one and rule-based one. For the latter, a really impressive amount of rules (in fact, thousands) is being developed and partial local programmes written on their basis are constantly applied in an incremental way. This bootstrapping and evolutionary method, in fact a whole bunch of methods, does yield a gradual success. As a consequence of this, there are now several versions of some corpora available, due to their different markup in this way and because of the need to preserve their referential status.

Not all the Czech problems in this are due to the specifically complicated morphology of the language, however. There are general problems in the field no one has satisfactorily and comprehensively solved so far in any language, though the situation varies here considerably, also due to the typological character of the language in question. Among the most serious problems is

handling of **multi-word forms**, completely or mostly neglected, usually. It is just not possible to be satisfied with mere text tokens where both grammatical forms (of complex verb tense forms of the kind *have seen*) and complex lexemes (idioms or terms, such as *take for granted* or *nitric acid*) have not been brought together by lemmatization and viewed as single lexemes and lemmas. There are many more idioms in everyday language use than one is aware of, this point being just where technicians and computational linguists do not mind while corpus linguists despair. It is very much a long-term task and programme to set this right. Another perennial problem, to name just one more, is distinguishing between **proper and common names**, not only because of a considerable overlap but also because of a constant influx of foreign surnames coming into one's language, usually through newspapers, and into the corpus. The new foreign names are nowhere to be found and their inflection is far from being always clear. This, too, is one of the major sources of the experience now well established anywhere, namely that the size of text forms recognized by a tagger and lemmatizer is still about fifty percent. The implication is that half of our enormous work is useless having no tag and lemma for this half. Fortunately, the data are still there and one can always query the forms there at least. Obviously, languages with less inflection might have less problems of this kind.

5. Methodology and Research

The study of language, made easier and more reliable by the existence of corpora now, has always been based on identification of **regularities**, **rules and relations** deduced from language data. It seems now, that the time for a dramatic change in linguistics has come, allowing one to study fully the **syntagmatic and combinatory aspects** of language for the first time and to redress the balance and past practice, which has always been bent on **paradigmatics**, categorizing, classification etc., without really having enough data for doing this. However, the obvious part played in this by the researcher's **introspection**, has never been critically questioned and it is only now that one is able to see to what degree this has been applied, contrary to facts one has now. The assumptions made by linguists and teachers about language on the basis of really few examples are staggering. Equally problematic are now past judgements about what is correct and what is wrong in language, a subject dear to hearts of all prescriptivists. There is no such thing as right or wrong, if it is not supported by data from real usage. The abyss between wishful thinking, recommended to and even imposed on the others in some cases, and real language facts is considerable. It is now clear that major revisions of old pre-conceived and unwarranted claims and made-up descriptions are due to come.

The language reality shown by corpora is no longer black and white with fine distinctions and simple truths built on these.

The new emphasis on syntagmatics, contrary to the paradigmatic past, requires new methodologies to be developed starting from the strategy of samples, ways how to handle clines and scales instead of the old pigeon-holing and marrying new statistical methods, these being still far from satisfactory. As these tasks that have to be solved in new pertinent and effective ways touch the very heart of linguistics, revision of all of the past will be necessary: there are first corpus-based grammars and dictionaries available now and these do tell a different story from the old bookshelf manuals. An answer starting with fresh insights is to be seen in corpus linguistics, as a new, unprejudiced study branch.

6. Current Shape of CNC

Aiming at a billion-word synchronous corpus by 2011 at least, the goal being almost attained by now, the current state of the whole Czech National Corpus Project can be described in its main sections as follows:

Type	**Corpus (+year)**	**Size**
Contemporary:		
Written synchronic:		
	SYN 2000	100 mil (balanced)
	SYN 2005	100 mil (balanced)
	SYNPUB 2006	300 mil (i.e. newspapers and magazines)
	SYNPUB2009	700 mil (newspapers and magazines)
Spoken synchronic:		
	PMK (Prague Spoken Corpus)	0.7 mil.
	BMK (Brno Spoken Corpus)	0.5 mil.
	ORAL 2006	1 mil (informal spoken language)
	ORAL 2008	1 mil (informal spoken language sociolinguistically balanced)

Historical:

DIA	1,8 mil (7 centuries from 13th cent.).

Specialized Corpora:

Karel Čapek' Corpus	2.6 mil.
Bohumil Hrabal' Corpus	1.7 mil.
Corpus of Totalitarian times	14 mil.
Hand-written letters KSK	0,8 mil (2000 letters)

Parallel Corpora (Czech vs 20 languages, in progress) manually aligned fiction, total size of the Czech part is currenly 25 mil. tokens

These are only the corpora that have already been published, but there is much more data in the Bank of CNC waiting for a further treatment and eventual release. On the other hand, the data here is constantly growing.

7. Outlook and Open Problems

Though there are, undoubtedly, new entities, phenomena and constructions to be yet discovered, certain things have already been recognized as new and important aspects of the corpus research. With the current, though often a bit fashionable, interest in discourse (whatever that means) a possibility to study language stereotypes and their use in authentic contexts has opened. With respect to their frequent use this will become a vast field of study (including phraseology). Another fact, new and uncomfortable for traditional linguists, is the corpus linguist's insisting on the study of word forms in contrast to the old interest in and description of lemmas only. It is evident that many word forms have a specific meaning bound to it only (eg. to a special personal form of a verb in one tense only, etc.) that are not to be found in the rest of its lexeme. The implication is to basically revise the existing dictionaries, among other things.

The Czech National Corpus project is an attempt at an unprecedented documentation and mapping of a language, reaching increasingly more from the present into the past and, hopefully, into the future as well. It will eventually enhance our understanding of the nature of the language across time, its major trends, core and periphery. As a supreme mirror of our reality and culture it offers an objective knowledge of facts, preventing unwarranted subjective conclusions, based on the introspection of a single man. Bridging different times in its development offers glimpses of the whole, but not only that. As no community and its language lives in isolation, knowledge and research of the

links and ties to other languages are necessary, too. A modest start of this may be seen in a rather recent subproject of InterCorp[3], a parallel corpus of some 20 languages linked to Czech, where also small languages such as Latvian, Lithuanian, Finnish, Dutch or Serbian are included. Obviously, to be able to go on, one must not resign and stop, having achieved a corpus of, say, a hundred million words. On the contrary, as the language is in constant flow and change serving the needs of our society that is not going to stop its development one day, it is necessary to upkeep its mapping in a corpus, too. Hence the need for a continuous corpus project and its support that must be firmly institutionalized and financially anchored.

It would be wrong to take a corpus for a general medicine for linguistics and all those disciplines dealing with the language. Today's corpus (let alone Internet) may not have answers for everything one might want to know, but it does offer at least hundred times more than the archives of recent past. Corpora have arrived and are likely to remain here whatever that might mean and imply. Realizing what they can offer and, in terms of human labour, what time they can spare us one must really appreciate the fact that we have them.

References

General

Burnard, L. (1995). *Users' Reference Guide for the British National Corpus*. Oxford U. Press, Oxford.

Burnard, L. (2007). Spoken Corpora Design Revisited
Available at: http://www.corpus.bham.ac.uk/corplingproceedings07/

Burnard, L. (2009). *Spoken Corpora Design: Their Constitutive Parametres*. IJCL 14: 113–123.

Burnard, L. (1995). "Jazykový korpus: Prostředek a zdroj poznání". *Slovo a slovesnost* 56: 119–140 (Language Corpus: Means and Source of Knowledge).

Burnard, L. (1997). "Czech National Corpus: A Case in Many Contexts". *International Journal of Corpus Linguistics* 2: 181–197.

Burnard, L. (1998). "Czech National Corpus: Its Character, Goal and Background". In: P. Sojka, V. Matoušek, K. Pala, I. Kopeček (eds.). *Text, Speech, Dialogue, Proceedings of the First Workshop on Text, Speech, Dialogue-TSD'98, Brno, Czech Republic, September*. Masaryk University: Brno: 9–14.

Burnard, L. (2001). "Language Corpora: The Czech Case" In: *Text, Speech and Dialogue, TSD 2001*. Eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer Springer Berlin etc.: 21–30.

Čermák, F., Králík, J. and K. Kučera (1997). "Recepce současné češtiny a reprezentativnost korpusu". *Slovo a slovesnost* 58: 117–124 (Reception of the Contemporary Czech and the Representativeness of Corpus).

3 http://ucnk.ff.cuni.cz/intercorp/

Český národní korpus. Úvod a příručka uživatele (2000). Eds. Kocek J., Kopřivová M., Kučera K., Filozofická fakulta KU Praha (Czech National Corpus. An Introduction and User's Manual).

Králík, J. and M. Šulc (2005). "The Representativeness of Czech Corpora". *IJCL*: 357–368.

Kruyt, J.G. (1993). *Design Criteria for Corpora Construction in the Framework of a European Corpora Network*. Final Report. Institute for Dutch Lexicology INL: Leiden.

Kučera, K. (1998). "Diachronní složka Českého národního korpusu: obecné zásady, kontext a současný stav". *Listy filologické* 121: 303–313 (Diachronic Component of the Czech National Corpus: General Principles, Context and Current State of Affairs).

Kučera, K. (2002). "The Czech National Corpus: Principles, Design, and Results". *Literary and Linguistic Computing*. Vol. 17, No. 2: 245–257.

Kučera, K. (2007). "Mapping the Time Continuum: A Major Raison d'être for Diachronic Corpora". In: M. Davies, P. Rayson, S. Hunston, and P. Danielsson (eds.). *Proceedings of the Corpus Linguistics Conference CL2007*. University of Birmingham, 2007, 10 s. http://www.corpus.bham.ac.uk/corplingproceedings07.html

The Design of the National Corpus of Polish

Rafał L. Górski

Abstract: The paper describes the proposed design of the National Corpus of Polish. The corpus shall reflect the perception of the language by the Polish linguistic community. Sources allowing for the reconstruction of the structure of readership in Poland as well the method leading to the final design of the are discussed.

Keywords: text, channel, corpus design, general-reference corpus, National Corpus of Polish, readership, representativeness, text types.

Corpora significantly differ in their design. This is due to two reasons. First of all there is no consent what is the reality which a corpus should represent. Further even if there was such a consent, each corpus represents a different linguistic community. In Górski (2008) we discuss several approaches to the question of the representativeness of a corpus. We suggested to adopt the concept of representation of the perception of the language by a certain linguistic community, as far as the written part of the corpus is concerned (cf. Čermák *et al*. 1997).

1. Balance and representativeness

Although the terms balance and representativeness are used interchangeably, we shall distinguish them for the purposes of the present paper. A *representative* corpus is a corpus which represents a certain reality (at this stage no matter which), a *balanced* one is a corpus which is not dominated by one text type. There is also a number of texts which are scarcely read, however cannot be omitted in a corpus if we expect it to reflect the entire language. These three requirements are often mutually exclusive. If we ask linguists for their preferences probably most of them would rather choose a balanced although not representative corpus than the other way round. Another problem, which never can be resolved is the proportion of the spoken component – measuring the right proportion is hardly feasible. Intuitively however we can say that an average representative of a linguistic community perceives much more spoken than written language, thus the proportions of these two channels in the corpus should favour the former.

Taking it all into account the design of the corpus is always a compromise between the strict requirements of adopted methodology, feasibility and usefulness of the compiled corpus.

2. Text typology

First prerequisite of any work on the design of a corpus is to set an appropriate typology of texts[1]. At the starting point we tried to elaborate a relatively simple and flat text typology, especially because assigning a given text to one of the classes is not always a straightforward task. The bigger granularity of the adopted typology the more difficult and prone to mistakes is the task. What is more we cannot say much about the readership of marginal text types as drama[2]. On the other hand some more detailed information (eg. the distinction between prose and poetry) is important to the end-user of the corpus. Again we tried to meet a compromise: texts are described by a more granular typology, however there is a certain hierarchy of categories: some of them consist of several sub-categories.

The set of text types consists of: fiction (with subcategories: prose, poetry and drama); non-fiction literature; journalism[3]; academic writing and textbooks; instructive writing and guidebooks; unclassified non-fiction book; miscellaneous (written) (with subcategories: legal and official; advertisements; announcements; political marketing; user manuals; letters); Internet (subcategories: interactive (forums, chat rooms, instant messaging, mailing lists); static WWW pages). Three further categories encompass various types of spoken texts: conversational; spoken from the media; quasi-spoken.

This typology requires some comments: it may seem that some texts are defined rather by their channel than strictly by the text type (journalism, Internet). Of course there is a number of literary or legal texts published in newspapers, still most of the content of press is journalism. On the other hand it is very difficult to filter out non-journalistic texts in press. The solution is to treat everything what is published in non-specialised press as journalism[4]. Again Internet is not a specific sort of texts but a channel. There are however texts

1 A detailed description the typology adopted for the corpus will be published elsewhere.

2 Marginal in means of audience not – say – importance for the culture.

3 By journalism we understand both short reports and opinion journalism, which are in fact two distinct (although close) text types. We decided to merge them because of lack of means of automatic sorting texts from newspapers.

4 Strictly speaking the press – with some exceptions – in the corpus is introduced via Internet so as to avoid digitalization of printed material. If one wants to be very precise the channel of these texts should be labeled as Internet. The divulgation of these texts primarily however disseminated as print.

which occur only on the Internet as e.g. dynamic www pages (blogs, forums etc.) thus in these and only these cases the channel defines the text type. It is worth noting that there is a certain hierarchy of text types. The less granular hierarchy is more essential. Thus we care more about the proportions of press, fiction, and non-fiction[5], then – inside the category of non-fiction – the proportions between academic writing and non-fictive literary work. We shall also stress that the more granular categories listed above in brackets serve only for a more precise description of text and play no role in representativeness.

The procedure of assigning a text to a particular class is as follows: what is published in general non-specialised press is assigned to the category journalism. In case of books first we distinguish fictive from non-fictive books. If the book is fictive – no further decisions have to be made. Otherwise the book has to be assigned to one of five remaining categories. Academic writing and textbooks form a clear category – an informative text written by specialist for specialists. The target distinguishes it from a similar category, namely instructive writing and guidebooks. This broad category encompasses: all kinds of instructive writing, as well as how-to books, tourist guides, cookbooks etc. Contrary to the former category its audience consists non-specialists. Another very broad category is non-fiction literature. It includes all non-fictive literary works, as well as biography, memoirs, non-fictive novels etc. The category is defined rather negatively, that is it is non-fictive on the one hand and but still shows some narration. Journalism is published not only in newspapers but also in books – and in principle we do not distinguish them from texts published in press. The last category – "unclassified non-fiction book" encompasses non-fictive books which either belong to a well defined category, however so marginal that it is not worth creating a different category, such as eg. collections of sermons, or belong to more than one category, or – last but not least – it is very hard to classify them. This typology is based basically on the Polish mainstream research on stylistics (cf. Gajda 1995, Klemensiewicz 1982).

The classification of the spoken component is based on different grounds. Note that we do not introduce one category "spoken", because all the three categories differ so much, that simply narrowing a query to it would be useless. Thus we distinguish transcripts of spontaneous dialogs from transcripts of texts spoken in media. The category quasi-spoken is created for texts which are primarily spoken, but are then edited and turned to a written text, as parliamentary transcripts or radio interviews published on the Internet. These texts while being written rather than spoken, still keep features of the spoken language. This means that there is a different rationale behind the typology of

5 Górski and Łaziński (forthcoming) prove that the distinction between fiction and non-fiction is also fundamental as far as intralinguistic features are concerned.

spoken and written texts. In case of spoken texts we take into account the setting (spontaneous vs. formal), but also the method of transcription of the text, or to put it in other words – the faithfulness of the transcript.

Apart from the above mentioned typology we classify texts by a channel with the categories: press (subcategories: daily; weekly; monthly; other press), book; Internet; spoken; leaflets, announcements, ads; manuscript. Every text is characterised by both text type and channel.

So as to assure that the corpus will cover as many topics as possible, we also include in headers the classification done by librarians: Universal Decimal Classification and the classification of the National Library (Biblioteka Narodowa). We shall stress however that we do not want the corpus to be representative according to topic. We use the classification as auxiliary means to check if the topics are covered evenly; it also helps in classifying texts according to the text typology of the corpus.

3. Methods

We decided to reflect in the corpus the reception of the language by the Polish language community (as far as the written component is concerned) rather then the population of texts. The motivation for this choice was discussed in detail in Górski 2008.

This methodology means in practice that we try to reflect the structure of the readership in Poland. To put it simply: the more words of a certain text type an average Pole reads, the larger is the proportion of this text type in the corpus.

3.1. Sources

Ideally we should conduct a pool asking the respondents about the volume of texts of different text types, that they read. However if we expected reliable data such a pool would be expensive and time consuming. Instead we can make use of data which are publicly available.

There are two main sources of the data in question. First: a biannual report of the National Library (Straus *et al*. 1996a, 1996b, 1998, 2000, 2002, 2006, 2008) A specialised department conducts a survey to answer the following questions: how many and what kind of books people read and buy. These surveys are managed by professional opinion poll agency and seem to be reliable. What is also important – the data are quite stable. Although in the nineties one could observe an increase of those who declare reading instructive

books, once their readership reached a certain point radical changes are no more observed.

As far as the press is concerned there are two sources available. The first is PBC (Polskie Badanie Czytelnictwa). It is a pool conducted twice a year and its aim is to show how many respondents declare reading a newspaper. In contrast ZKDP (Związek Kontroli Dystrybucji Prasy) controls the circulation of the press. In the latter case we can assume that everybody who bought a newspaper read it. It is not necessarely true in two cases – some people accept free press, just because it is free but do not read it and on the other hand a copy of a teenagers magazine is often read by a couple of readers. This however should not affect very highly the data. On the other hand ZDKP controls a wider range of magazines and for that reason we chose it as a basis for establishing the design of the corpus.

3.2. Methods

We adopted a bottom-up method of establishing the design of the corpus. At the first step we decide about the proportions of books. As we wrote above ideally we should know how many running words did an average Pole read in books belonging to a certain text-type. Instead, the quoted above reports of the National Library tell us how many respondents declare reading a certain type of texts. In fact it is not exactly what we want. Still however it tells us about the audience of a given text type. The larger is the audience the wider is the reception of this text type. Thus what we do is we assign the volume of texts belonging to a certain text type proportionally to the percentage of respondents declaring reading this very text type. Now, we count the average of the percentage of respondents declaring reading each text type in the available reports and normalize it so as to obtain the percentage of certain text types characteristic for books.

The figures in the following table show mean percentage of respondents declaring reading a given text sort. The first column contains Polish labels of text categories the second their English translation[6], and the third the mean percentage of respondents declaring reading this text type. Each respondent could choose as many text types as he wants, so the figures do not sum up to 100. Note that the typology set by the poll differs form the one which we propose. It is however possible to "translate" one to another.

6 One has to be aware of the fact that the English counterparts of the Polish terms do not match exactly. For example Polish *esej* has a narrower meaning than its English counterpart essey – it is a kind of literary work, thus it excludes scientific writing.

	Type (Polish)	Type (English)	Mean percentage of respondents
1.	lektury szkolne i podręczniki	literary set books and textbooks	23%
2.	obyczajowo-romansowe	romantic stories	18%
3.	sensacyjno-kryminalne	detective stories	16%
4.	dziecięco-młodzieżowe	novels for youth and children	10%
5.	encyklopedyczno-poradnikowe	encyclopaedias and how-to books	14%
6.	literatura faktu (wspomnienia, pamiętniki, reportaże, biografie, autobiografie)	non-fiction literary works (memoirs, reportages, biography, authobiography)	13%
7.	powieści grozy, horrory, thrillery	horrors and thrillers	3%
8.	fantastyka (science-fiction, fantasy, litertatura grozy z horrorami)	fantasy	6%
9.	fachowe	professional literature	8%
10.	religijne	religious writing	8%
11.	eseistyka i publicystyka	esseys and opinion	3%
12.	ezoteryka i ufologia	esoterics and ufology	2%

As you can see these categories describe rather the "literary taste" or cultural needs of respondents than text sorts. Especially the row 1 is unclear: it may cover any kind of book which the respondent was forced to read by the school, that is why we do not take this figure into account. If we add rows: 2, 3 4, 7, 8, (which are all various genders of belles-lettres) and treat books as 100% we obtain following figures:

Gender	percentage
fiction	52%
non-fiction literature	19%
journalistic book	5%
academic writing and textbooks	5%
instructive writing and guidebooks	19%

Now let us turn to newspapers. Note however that there are two good reasons to treat books and newspapers separately at this stage: we have quite distinct data, which do not match each other, according to those two channels of text. As far as newspapers are concerned we only know how many copies of each title are bought. The proposed typology of press is a very simple one – we distinguish dailies and magazines. Although there are detailed typologies of press established within media studies, they are useless, because almost every title belongs to several categories at once. On the other hand this rather oversimplified classification has a strong justification. Dailies contain both

informative writing and persuasive texts (the former outnumbering the latter), whereas magazines contain predominantly the latter. In fact short news and opinion journalism form two distinct text types. Now, we count the sum of all copies of magazines in a year and again the same for dailies. The proportion between these two sums is the proportion which we obtain for these two channels, that is 52% dailies and 48% magazines.

Having already established the design of the two main components of published texts we pass to the most delicate question of the proportions of books and newspapers. The book readership surveys tell us that an average Pole reads 8 books a year. Assuming that the length of an average book is 70 000 running words (a figure suggested by the corpus), the annual "consumption" of texts read in books is 560 000. Unfortunately the precise amount of text read in newspapers is not monitored so accurately[7]. There is one publicly available source (Makarenko 2001) which states that an average Pole reads the press ca. 25 minutes a day. With the average speed of 200–230 per minute this gives us ca. 1 960 000 words a year (215 x 25 x 365 = 1 961 875). If so, the percentage of words read in press is ca. 78% against 22% read in books.

As stated above, we would like the corpus to *represent* the entire language in its variety. Thus we add a component labelled "other" or "miscellanea". In case of this component we only care about diversity so as to cover all possible kinds of texts. No a priori set amounts of texts are foreseen. In fact it is impossible to state the audience of such texts and probably some of them are hardly ever read (e.g. legal texts), other attract broad public (internet). Thus we arbitrarily assign them a share of 10% of the entire corpus, which breaks down into 7% Internet texts and 3% of "miscellaneus (written)". The other arbitrarily set part of the corpus is the spoken component which is also covers 10% of the corpus.

If we put all of it together we would obtain:

press	62%
books	18%
miscellanea	10%
spoken	10%

4. Final design

Recall however that we assume balance as an equally important feature of corpus design as representativeness. Let us define a balanced corpus as a corpus in which no component forms more than the halve of it. Again we arbitrarily

7 This for an obvious reason – the advertiser cares only about reading a newspaper, what means to him reading also his advertisement. The question how much of the newspaper is really read is of no importance to him.

decide to lower the amount of journalistic texts to 50%. There is also another motivation for this change – in general newspapers are read less carefully than books[8].

Taking this assumption we finally propose following design of the corpus:

journalism	50%
fiction	16%
spoken	10%
internet	7%
non-fiction literature	5.5%
instructive writing and guidebooks	5.5%
miscellaneus (written)	3%
academic writing and textbooks	2%
unclassified non-fiction book	1%

In case of a large corpus there is no need of extracting samples from the texts, because no single text can skew the data. What is the more one should not "waste" text if the corpus is to be very large. Although a random selection of texts should be desired (cf. Biber 1993) it never happens in modern corpora. Each text to become part of a corpus must satisfy two conditions it should be available in electronic form and the agreement of a copyright holder must be obtained. Of course there is a danger of self-selection, for example publishers are not prone to give their permission for best-sellers. This is however the cost of building a large corpus.

References

Biber, D. (1993). "Representativeness in Corpus Design". *Literary and Linguistic Computing*, 8 (4): 243–257.

Čermák, F., Králík, J. and K. Kučera (1997). "Recepce současné češtiny a reprezentativnost korpusu". *Slovo a Slovesnost*, 58: 117–124.

Gajda, S. (ed.) (1995). *Przewodnik po stylistyce polskiej*. Opole: Wydawnictwo Uniwersytetu Opolskiego.

Górski, R. L. (2009). "Representativness of the written part of a Polish general reference corpus. Primary notes". In: B. Lewandowska-Tomaszczyk (ed.). *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. Frankfurt a. Main: Peter Lang: 119–123.

Górski, R. L. and M. Łaziński (forthcoming). „Wzór stylu i wzór na styl. Zróżnicowanie stylistyczne tekstów w Narodowym Korpusie Języka Polskiego". To appear in: VIII Forum Kultury Słowa, Gdańsk.

8 This fact was pointed to me by František Čermak (personal communication). Even if I do not know research on this topic I suggest a simple test – most people remember both the plot of a novel and even some quotations from them but hardly can tell that of a daily read a week ago.

Klemensiewicz, Z (1982). „O różnych odmianach współczesnej polszczyzny". In: Z. Klemensiewicz (ed.). *Składnia, stylistyka, pedagogika językowa*. Warszawa: Państwowe Wydawnictwo Naukowe.

Makarenko, V. (2001). „Nie tylko liczyć, trzeba też rozumieć". *Gazeta Wyborcza*, no 288: 22 Available at: http://szukaj.wyborcza.pl/archiwum/1,0,1599476.html

Straus G. and K. Wolff (1996a). *Czytanie i kupowanie książek w Polsce w 1994 r.* Straus G. and K. Wolff (1996b). *Polacy i książki: społeczna sytuacja książki w Polsce 1992*. Warszawa: Biblioteka Narodowa.

Straus G. and K. Wolff (1998). *Zainteresowanie książką w społeczeństwie polskim w 1996 r*. Warszawa: Biblioteka Narodowa.

Straus G. and K. Wolff (2000). *Czytać, nie czytać..., kupować, nie kupować... sytuacja książki w społeczeństwie polskim w 1998 r.* Warszawa: Biblioteka Narodowa.

Straus G. and K. Wolff (2002). *Sienkiewicz, Mickiewicz, Biblia, harlequiny...: społeczny zasięg książki w Polsce w 2000 roku*.Warszawa: Biblioteka Narodowa.

Straus G. and K. Wolff (2006). *Czytanie, kupowanie, wypożyczanie: społeczny zasięg książki w Polsce w 2004 roku.* Warszawa: Biblioteka Narodowa.

Straus G., K. Wolff, S. Wierny (2008). *Czytanie, kupowanie, surfowanie: społeczny zasięg książki w Polsce w 2006 roku.* Warszawa: Biblioteka Narodowa.

Polskie Badanie Czytelnictwa www.pbczyt.pl

Związek Kontroli Dystrybucji Prasy www.zkdp.pl

XML Text Interchange Format in the National Corpus of Polish

Adam Przepiórkowski and Piotr Bański

Abstract: The aim of this paper is to describe and justify the XML encoding of texts within the National Corpus of Polish. Basic text encoding, rather than linguistic annotation, is considered here: the encoding of the primary data, the structural markup and the metadata. A set of schemata conformant with the Text Encoding Initiative Guidelines P5 is presented.

Keywords: Text Encoding Initiative, TEI P5, National Corpus of Polish, NKJP, metadata, TEI header, structural markup, primary data, XML, Polish.

1. Introduction

National Corpus of Polish[1] (Pol. *Narodowy Korpus Języka Polskiego;* NKJP; http://nkjp.pl/) is a project carried out in 2008-2010, involving 4 Polish institutions: Institute of Computer Science of the Polish Academy of Sciences (coordinator), Institute of Polish Language of the Polish Academy of Sciences, University of Lódź and Polish Scientific Publishers PWN.[2] Each of these instituti-tions contributes texts from their own corpora, and each – apart from the coordinator – acquires new texts for the National Corpus of Polish (NKJP, henceforth): books, newspapers and magazines, blogs, transcripts of spoken data, etc. All these texts are imported into two very different search engines available in NKJP (cf. the "Demo" link at http://nkjp.pl/).

Obviously, before NKJP texts can be indexed or automatically processed by any other tools they must be converted to a common interchange format. Such interchange format should allow for the representation of various types of texts mentioned above, and also for the encoding of various kinds of metadata and structural information. The only text encoding standard sufficiently versatile to meet these requirements is TEI P5, presented in the Guidelines of the Text Encoding Initiative (TEI; Burnard and Bauman 2008; http://www.tei-c.org/). It is not an official ISO standard, but a mature and very specific XML-based *de facto* standard for text encoding in the humanities, with a rich user base and supporting tools.

1 Research funded in 2007-2010 by a research and development grant from the Polish Ministry of Science and Higher Education.

2 A programmatic description of the project may be found in Przepiórkowski et al. 2008, and more recent developments are presented in Przepiórkowski et al. 2009.

The reason for continuing this paper beyond the previous paragraph is that TEI is a large treasure trove of solutions, rather than a lean and highly focussed formalism, and a particular text encoding schema must still be designed by choosing the most appropriate mechanisms from the TEI toolbox and – in rare specific cases – by introducing new XML elements or attributes. The aim of this paper is to present and document one such particular schema, developed within NKJP. As there are few well-documented TEI corpora around, and hardly any corpora following the current P5 version of the TEI Guidelines (substantially differing from the previous TEI versions), we hope that this presentation will facilitate the development of other TEI P5 corpora.

The remainder of the paper starts, in § 2, with a presentation of the NKJP corpus header, i.e., an XML document containing metadata pertaining to the National Corpus of Polish as a whole. The representation of text headers, i.e., metadata for particular texts, is described in § 3. The ensuing section, § 4, makes clear the overall structure of a corpus text and the place of both kinds of metadata in that structure. This section also sketches the representation of structural and typographical distinctions within texts. Although, within NKJP, texts are also annotated at various linguistic levels, this paper does not deal with such linguistic annotation – see Przepiórkowski and Bański 2009 for an overview and Bański and Przepiórkowski 2009 for a discussion of some technical issues. Finally, § 5 concludes the paper.

2. Corpus Header

Following the TEI Guidelines, the NKJP corpus header consists of 4 sections contained in the `<teiHeader xml:lang="en" type="corpus">` element: `<fileDesc>`, `<profileDesc>`, `<encodingDesc>` and `<revisionDesc>`.

Two of these have very simple structure. First, `<profileDesc>` identifies the main languages used in the TEI encoding of texts and metadata, and it is cited in its entirety below:

```
<profileDesc>
   <langUsage>
      <language ident="pl">Polish</language>
      <language ident="en">English</language>
   </langUsage>
</profileDesc>
```

The values of `@ident` attributes may be used for any element to specify the language of the content of that element. In fact, the `xml:lang="en"`

specification in the <teiHeader> element is inherited by other elements in the header, unless explicitly overridden by xml:lang="pl", thus making English the default language of the NKJP header.

Another simple and homogeneous section is <revisionDesc>: it contains a sequence of <change> statements like the following:

```
<change who="#adamp" when="2009-08-01">
   Added <gi>profileDesc</gi>.
</change>
```

The <fileDesc> section contains 4 subsections. The first, <titleStmt>, specifies the name of the corpus and describes the responsibility of various institutions and persons involved in its creation. One such responsibility statement is referenced by who="#adamp" in the example above, another may look as follows:

```
<respStmt>
   <persName xml:id="bansp">Piotr Bański</persName>
   <resp>initial design of various XML schemata</resp>
</respStmt>
```

The other three subsections of <fileDesc> are: <editionStmt> – a brief statement concerning the stability of the current version of NKJP, <publicationStmt> – defining availability and distribution of NKJP, and <sourceDesc> – specifying the origin of texts in general terms (specific source descriptions are contained in the headers of particular texts).

Finally, <encodingDesc> characterizes NKJP in various ways, e.g., <projectDesc> repeats the description of the project given at http://nkjp.pl/, <samplingDecl> says that *Whole texts are included, whenever possible* and provides some information on text structure, as discussed in § 4, and <editorialDecl> briefly discusses anonymisation of spoken data and other editorial interventions in NKJP texts.

While these subsections contain free-text statements, many other <encodingDesc> subsections are more structured. Perhaps the most important are <classDecl> subsections, which specify text classifications referenced in particular text headers. For example, one of the ways in which NKJP texts are classified is according to the Universal Decimal Classification, so the following declaration is present in the corpus header:

```
<classDecl>
   <taxonomy xml:id="ukd">
      <bibl>
         <title xml:lang="pl">Uniwersalna Klasyfikacja
         Dziesietna</title>
         <title xml:lang="en">Universal Decimal
         Classification</title>
         <edition>UDC-P058</edition>
      </bibl>
   </taxonomy>
</classDecl>
```

Within a text header (cf. § 3 below), a reference to this classification may be made as follows:

```
<classCode scheme="#ukd">821.162.1-3</classCode>.
```

Similarly, in order to control the good balance of the corpus with respect to genres, a taxonomy of text types is defined; its fragment is presented below:

```
<classDecl>
   <taxonomy xml:id="taxonomy-NKJP-type">
    <! -- ... --->
      <category xml:id="typ_lit_proza">
         <desc xml:lang="pl">proza</desc>
         <desc xml:lang="en">prose</desc>
      </category>
      <category xml:id="typ_lit_poezja">
         <desc xml:lang="pl">poezja</desc>
         <desc xml:lang="en">poetry</desc>
      </category>
      <category xml:id="typ_lit_dramat">
         <desc xml:lang="pl">dramat</desc>
         <desc xml:lang="en">drama</desc>
      </category>
    <! -- ... --->
   </taxonomy>
</classDecl>
```

Again, the type of a particular text may be defined by referencing one of the categories defined in such a classification.