Input a Word, Analyze the World

Input a Word, Analyze the World:

Selected Approaches to Corpus Linguistics

Edited by
Francisco Alonso Almeida,
Ivalla Ortega Barrera,
Elena Quintana Toledo
and Margarita E. Sánchez Cuervo

Input a Word, Analyze the World: Selected Approaches
to Corpus Linguistics

Edited by Francisco Alonso Almeida, Ivalla Ortega Barrera,
Elena Quintana Toledo and Margarita E. Sánchez Cuervo

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-8513-4
ISBN (13): 978-1-4438-8513-3

TABLE OF CONTENTS

INTRODUCTION

INPUT A WORD: ANALYZE THE WORLD

FRANCISCO ALONSO ALMEIDA,
IVALLA ORTEGA BARRERA,
ELENA QUINTANA TOLEDO,
AND MARGARITA E. SÁNCHEZ CUERVO

This volume presents current perspectives on corpus linguistics (CL) from a variety of linguistic subdisciplines. CL has proven an excellent methodology for the study of language variation and change, and is well suited for interdisciplinary collaboration, as shown by the studies in this book. The title of this monograph emerges precisely from the use of CL to assess language in different registers and with a variety of purposes. This collection contains 30 contributions by scholars in the field worldwide. These papers are organized into four parts, namely (a) corpus production and corpus tools, (b) lexical analysis, phraseology and grammar, (c) translation and contrastive linguistics, and (d) language learning.

Katarzyna Klessa and Tomasz Wicherkiewicz give an account of the design and implementation of an online database for the compilation of Poland's endangered language varieties. Antonio Moreno Sandoval, Leonardo Campillos Llanos, Carlos Herrero Zorita, José María Guirao Miras, Alicia González Martínez, Doaa Samy, and Emi Takamori present the MultiMedica corpus, a freely available online interface that includes biomedical texts in Spanish, Japanese and Arabic. The contribution by Sardar Jaf and Allan Ramsay focuses on the development of a new part-of-speech (POS) tagger for Sorani Kurdish. They use a rule-based method for building a corpus of Kurdish, a language where the only available resources are raw texts and a few descriptive grammar text books. The corpus will be used to develop an advanced POS-tagging tool. Begoña Crespo and Isabel de la Cruz Cabanillas describe the use of corpus

linguistics for historical linguistics purposes. The authors give an overview of selected historical corpora to exemplify computerized resources for historical studies of English.

Virginia Gutiérrez Rodríguez, Octavio Santana Suárez, José R. Pérez Aguiar, and Isabel Sánchez-Berriel describe ARGCoS (Automatic Recognition and Generation Compounds in Spanish), a useful tool to identify groups of simple forms that function as compound lexical units and also the combination of simple and compound forms to produce new compounds in Spanish. Wiebke Petersen and Oliver Hellwig offer *Aṣṭādhyāyī* 2.0, a project that develops a digital edition of Pāṇini's more than 2000-year-old grammar of Sanskrit. This digital edition, which includes the texts with annotations on all text components, can be accessed by a web interface that allows users to browse the text and search for a particular grammatical pattern. Svetlana Toldova, Yulia Grishina, Alina Ladygina, Maria Vasilyeva, Galina Sim and Ilya Azerkovich explain the design of the first Russian open-source text corpus annotated with coreference relations. Marcus Vinicius Carvalho Guelpeli and Heider Moura Fernandes describe the creation of a corpus that comprises texts belonging to different registers (journalistic, law and medical texts) in Spanish that have been excerpted from the Internet.

Jean-Pierre Colson's chapter deals with computational phraseology as the study of set phrases that cover a wide spectrum of n-grams. In particular, it details the results of a methodology that allows the extraction not only of common collocations but also of compound terms or clichés related to a specific word. Sultan Nasser Al-Mujaiwel's study applies corpus linguistics to the information organization of items in 24 Arabic lexicon corpora, analyzed through existing lexicographic Arabic works and KACSTAC, in order to test specific morphological forms of items and their collocations indicators. Suet-Ching Soon and Siaw-Fong Chung examine several Chinese locative particles that mean "on/above" like *shàng*, *shàngmiàn*, *shángtou* and *shàngbian*. They conclude that these particles occur with different elements that may have either an abstract or a concrete usage. Ying Liu, Naixing Wei and Alex Chengyu Fang's paper is concerned with a corpus-based report on the use of nominalizations across China English and British English in two comparable media corpora. Their results show that the languages prefer different stylistic choices depending on the type of informal or formal media categories used.

The chapter by Isabel Sánchez-Berriel, Octavio Santana Suárez, José R. Pérez Aguiar, and Virginia Gutiérrez Rodríguez investigates whether there is a preferable combination used to distinguish between collocations

and free combinations in a large Spanish corpus. Oriana Mendoza's chapter seeks to identify the relationship between young people and religion. By means of discourse analysis and corpus linguistics, the main objective is to understand how a Mexican community expresses its religiosity. Carleen Gruntman examines a possible single meaning of the polysemous English preposition *for*. By using the Corpus of Contemporary American English, the author demonstrates that a highly schematic meaning can be distinguished, and supports the existence of schematicity in human language. Luis Miguel Puente Castelo describes the uses of conditionals in the works of English-speaking writers on philosophy and life sciences in the eighteenth century. Through the Coruña English Tools, the article searches for evidence of a gender-motivated difference between the uses of conditional structures in these texts.

Stéphane Patin proposes the study of the lexical specificities of the debates on the state of the nation used by the former presidents José María Aznar and José Luis Rodríguez Zapatero during their governments. Patin uses a lexicometric approach to contrast the specific themes in a speech or group of speeches. The creation of an Arabic corpus for sentiment analysis is described in Saud S. Alotaibi and Charles Anderson's chapter. This corpus was collected from different websites and created for predicting subjectivity and polarity in documents and sentences.

Claudio Fantinuoli analyzes the differences between texts translated by computer-assisted translation (CAT) tools and texts translated without them. He focuses on four features, namely, (i) sentence length, (ii) lexical density, (iii) demonstrative pronouns, and (iv) nominal/verbal ratio. The results of this analysis have major implications for the design and compilation of corpora aimed at carrying out translation studies, as well as for the design of CAT tools themselves. Ramón Martí Solano's contribution offers a contrastive analysis of some adjective + noun collocations in English and their translations into French. The author shows that collocational translatability involving cognate adjectives in English and French is possible as they share one or more senses, allowing them to collocate with the same paradigm of nouns in both languages.

Carmen Luján García presents a corpus-based study of the use of proper names of Anglo-American origin in Spain. She compares the frequency of use of those proper names in a time span of two decades – 1960-1969 and 2000-2009 – in order to highlight the increasing impact of the English language and culture on Spanish society. Margarita Mele Marrero describes the use of the English – 's in Spanish with the aid of a substitute web corpus. She analyzes parameters like (i) integrity, (ii) paradigmaticity, (iii) paradigmatic variability, (iv) structural scope, (v)

bondedness, and (vi) syntagmatic variability. They all seem to point to a lexicalization of the English genitive that appears to facilitate its infiltration into other languages.

The chapter by María Goretti García Morales contains a study of anglicisms in Spanish film reviews. She analyzes specialized, semi-specialized and non-specialized sources, and finds that loanwords with various degrees of assimilation do not only occur in texts aimed at experts in the film industry but also in texts meant for the general public. María Teresa Ortego Antón discusses the treatment given to twenty units of specialized meaning in the field of IT in three general bilingual dictionaries. She points out that translators have several difficulties, mainly because of a lack of systematized labelling for these units and because no contextual information is provided. Philippe Humblé employs corpus linguistics tools in his examination of the English and Spanish translations of Kader Abdolah's novel *Spijkerschrift*. His results show that there has been variations in the type/token relationship, the number and size of clauses as well as the use of keywords with respect to the original text written in Dutch, rendering translations that are closer to the native norm in the target languages than to the "unregularised character" Abdolah actually wanted to transmit.

An Vande Casteele and Kim Collewaert study referring expressions in Spanish oral narratives in a learner corpus. The authors refer to the phenomena of overspecification and overexplication to account for the redundant use of pronouns and proper nouns. The chapter by Budi Irmawati, Mamoru Komachi and Yuji Matsumoto deals with the development of a corpus of second-language learners of Indonesian with native corrections. The authors explain their work in this respect and the steps to be taken in improving their corpus. The use of the corpus for language learning and teaching is also described.

The contribution by Katsunori Kotani, Takehiko Yoshimi and Mayumi Uchida offers a description of their learner translation corpus. This corpus was compiled by translations by EFL learners, and the authors analyze it to validate its usefulness. Miharu Fuyuno, Yuko Yamashita and Yoshitaka Nakajima focus on a multimodal corpus of Japanese ELF learners to analyze audio and video performances. Their research seeks to provide more information on the speaking aspect for EFL learners, including facial expressions. The results of their contribution will have implications for teaching speaking competences.

Finally, María Luisa Carrió Pastor describes the use of hedges in academic papers written by English, Spanish and Chinese scholars. Her study explores variations in the use of these rhetorical strategies to see

whether the writer's cultural constraints have an effect on the selection and use of hedging devices in the construction of new knowledge. Her study has clear didactic implications since awareness of multicultural variation in the use of hedges may lead to more effective teaching and learning of these devices.[1]

[1] The papers included in this volume are much-elaborated versions of the papers presented at the 6th International Conference on Corpus Linguistics held at Las Palmas de Gran Canaria (Spain), 22-24 May, 2014 (AELINCO). We would like to acknowledge the help of the Departamento de Filología Moderna, Universidad de Las Palmas de Gran Canaria, during the celebration of this International Conference.

PART I:

LEXICAL ANALYSIS, PHRASEOLOGY AND GRAMMAR

CHAPTER ONE

DESIGN AND IMPLEMENTATION OF AN ONLINE DATABASE FOR ENDANGERED LANGUAGES: MULTILINGUAL HERITAGE OF POLAND

KATARZYNA KLESSA[1]
AND TOMASZ WICHERKIEWICZ[2]

1. Introduction

The topic of language diversity and its steady disappearance is being increasingly studied, discussed and researched on various levels and from different standpoints. Generally, though, the notions of language endangerment and language extinction have been associated with far-off areas such as the islands of Oceania, a number of regions in Africa, the two continents of the Americas, Siberia and Australia (which represents almost every continent). Many linguists realize or acknowledge the accelerating processes of disappearance and endangerment of languages because they constitute an imminent risk to the discipline of linguistics itself. As Krauss (1992: 10) suggests:

> Obviously we must do some serious rethinking of our priorities, lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated.

The diminishing diversity of languages of the world is steadily becoming an issue of public discourse in mass media and political fora. In that

[1] Institute of Linguistics, Faculty of Modern Languages and Literatures, Adam Mickiewicz University in Poznań. Email: klessa@amu.edu.pl.
[2] Department for Language Policy and Minority Studies, Faculty of Modern Languages and Literatures, Adam Mickiewicz University in Poznań. Email: wicher@amu.edu.pl.

respect, the European media and society tend to promote and give the impression that actually more languages are appearing than disappearing in the European language repertoire. The political developments at the turn of the 20th century and the resulting language engineering practices in the Balkans and post-Soviet regions, on one hand, and the European language policy aimed at protecting and promoting linguistic diversity, on the other, have also contributed to a growing consciousness of language diversity issues, including the methods and objectives of language preservation and planning (cf. Nau and Wicherkiewicz 2013). The topic of 'Endangered Languages' is progressively becoming an autonomous research/academic and pragmatic/political issue (cf. e.g., Austin and Sallabank 2011) while 'Minority Languages' and 'Language Diversity' tend to be among the most frequently referred to issues of language policy worldwide.

The position of the endangered languages issue in the European debate has not prevented the continent's linguistic inventory suffering from the global tendency of language disappearance (mainly through the diminishing intergenerational transmission of lesser-used minority, regional and non-territorial languages). The *Atlas of the World's Languages in Danger,*[3] the former *Red Book of Endangered Languages,* has repeatedly referred to extinct languages such as Slovincian in Poland; critically endangered ones such as Karaim in Ukraine, Livonian in Latvia (actually already extinct as a first language), Cornish and Manx (successfully revitalized Celtic languages); severely endangered languages such as Kashubian, Wilamowicean (Vilamovian), Karaim in Lithuania, and Saterlandic (Saterfrisian) in Germany; definitely endangered such as Romani in many countries, Lower Sorbian in Germany and even Irish in Ireland; vulnerable languages (e.g., Upper Sorbian, Belarusian, Rusyn); and many other languages/dialects in danger of extinction.

As outlined by Wicherkiewicz (2014), the territory of Poland has always been inhabited by numerous communities speaking languages other than Polish. Naturally, the communities who spoke these languages changed, reducing them to means of communication and tokens of (objective and subjective) identity. Throughout history, the territory of *Rzeczpospolita*[4] in particular has changed more than any other political entity in Europe (see also Figure 1).

[3] http://www.unesco.org/culture/languages-atlas

[4] *Rzeczpospolita* (calque from Latin *res publica*) is the Polish endonymic term referring to the consecutive state(hood) forms: Commonwealth of Both Nations (*I Rzeczpospolita*), the interbellum independent Polish state (*II Rzeczpospolita* – 1918-1945), the People's Republic of Poland (1945/1952-1989), and the *III Rzeczpospolita* (Since 1990).

The multinational character of the First Republic (up to 1795) and the Second Republic of Poland (1918-1939) stemmed from the vastness and diversity of the lands, which once constituted the largest country in Europe. To mention all vernaculars spoken at that time on the territory of the Commonwealth and its fiefdoms, one would have to take account of all Western[5] and Eastern Slavic languages, all Baltic languages,[6] some Fennic (south Estonian and Livonian) varieties, Low- and Middle-German vernaculars, and Yiddish as well as the Turkic languages of the Karaims, Tatars and Armenians.

It was already during the Second Republic that Poland's linguistic landscape was becoming restricted. Nonetheless, in the interwar period, Poland was still a country of many ethnicities, religions and languages. Some of these languages were spoken by communities numbering hundreds of thousands, as the results of the 1931 population census showed. The population of Poland (which, in 1931, was 31.9 million people) was (according to the declaration of one's native language) 69% Polish-speaking. Other languages spoken by Poland's citizens included Ukrainian (10.1%), Ruthenian[7] (3.8%), Jewish (=Yiddish 7.8%), Belarusian (3.1%), *Local* (2.2%),[8] German (2.3%), Hebrew (0.08%), Russian (0.04%), Lithuanian (0.03%), and Czech (0.01%). Accordingly, nearly 10 million citizens of interwar Poland spoke not Polish but a different language on a daily basis.

Following the end of World War II, the citizens of (the People's Republic of) Poland found themselves in a completely different reality, a new linguistic one included. The borders were moved westwards – the Germans were expelled to Germany, the Ukrainians and Rusyn Lemkos to the Soviet Union and to the "recovered" territories in Western and Northern Poland. The Jewish and Gypsy peoples had fallen prey to the Nazi German murder machine. All this, along with massive external and internal migrations and the new policies taken up by the Polish communist government in cooperation with the USSR, led to a linguistic and national unification. Minorities were no longer discussed or written about.

---

[5] Including Pomeranian varietes (ancestor of Kashubian)

[6] Including extinct Prussian and Yatvingian

[7] What draws attention is the 1.2 million speakers of "Ruthenian" (*ruski*). At that time, this was the term employed in relation to speakers of Ukrainian (especially in the formerly Austrian part of Poland) – thus, to sum up, users of Ukrainian accounted for nearly 14 per cent (4.4 million) of the citizens of the II Republic of Poland.

[8] The "local" language (*tutejszy*) was declared mostly by the inhabitants of Polesie, a district situated at the linguistic Belarusian-Ukrainian border.

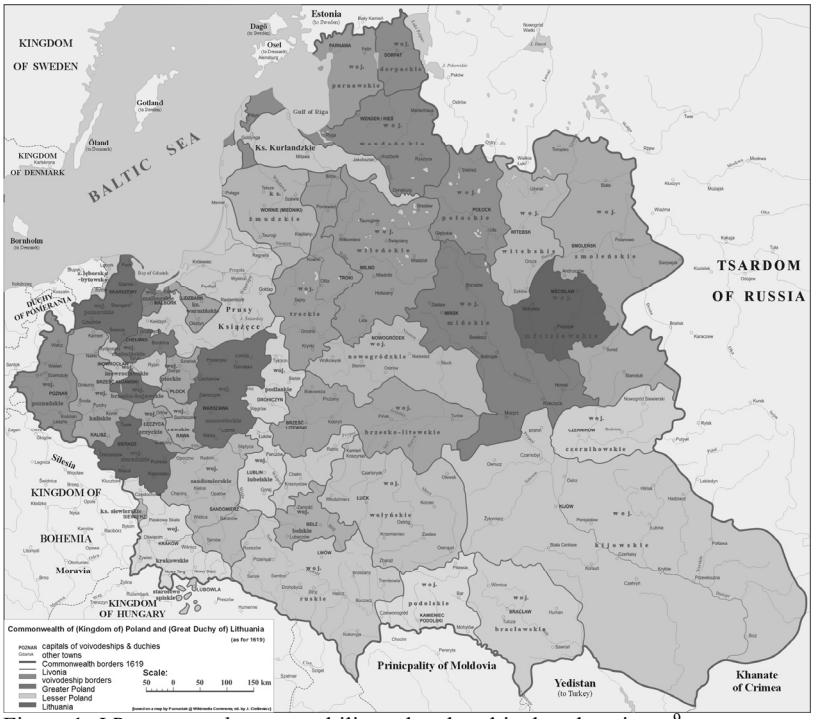

Figure 1. *I Rzeczpospolita* – a multilingual and multicultural territory.[9]

Only after the changes of 1956 were the minority and language policies mitigated. Schools offering the teaching of standardized varieties of minority languages were founded – German, Jewish, Ukrainian, Belarusian, Lithuanian, Slovak, Czech, and even Greek (for children of political refugees from Greece). They operated with newly published course books. Magazines in German, Yiddish and Hebrew, Ukrainian, Belarusian, Russian, Lithuanian, Czech, and Slovak or Romani were published. The government even allowed Kashubian and Lemko publications, even though, at the time, they were deemed mere dialects of Polish and Ukrainian, respectively. A Greek magazine with a supplement in Macedonian was also published for the refugees from dictatorial Greece. In different periods, radio programmes were broadcast in Belarusian, German, Lithuanian, Russian, Ukrainian, and Yiddish. In this

[9] The map displayed in Figure 1 is one of a number of maps created for the present project (full colour version: http://innejezyki.amu.edu.pl/Frontend/ Home/About).

new environment, the already small and dispersed minorities were assimilated at an ever-greater pace under the constant control of the government and without any consistent language policy to protect them. During the communist period, no statistics for ethnic or language minorities living in Poland were published or even known, and the recognition of existing minorities was of a marginal and selective nature. The official ethnic and language policy was characterized by intermittent periods of limited concessions granted to state-controlled organizations of some minorities (especially the so-called "socio-cultural societies" established after the thaw of 1956 for Belarusians, Czechs and Slovaks, Jews, Lithuanians, Roma, Russians, and Ukrainians). The public presence of the minorities and their languages, however, was limited to mostly folklore aspects. The very existence of other groups, such as Germans or Rusyn Lemkos (accepted exclusively as a sub-group of Ukrainians), was either ignored or overtly denied. The Tatars, Karaims and Armenians were treated principally as exotic religious minorities (for more details of minority and minority language policy and developments in communist Poland, see Majewicz and Wicherkiewicz 1990; 1998).

Transformations of the system, which started in 1989, brought an immediate change in the state policy towards minority issues, followed by prevailingly positive societal support expressed in public debates on various levels and a dramatically growing interest in problems of ethnicity and minorities. The previously recognized groups (Belarusians, Czechs and Slovaks, Jews, Lithuanians, Roma, Russians, Ukrainians, Tatars, and Karaims) as well as Germans and Armenians gradually adapted to the new conditions and started reformulating their (language) policies, even if originally to a very limited extent. The process of official and public recognition of the German and Rusyn-Lemko minorities proved the most difficult: their existence and official representative structures were eventually acknowledged only at the beginning of the 1990s.

In order to specify and legally secure political provisions, in 2005 the Polish Parliament passed a *law on national and ethnic minorities and on the regional languages*, officially recognizing nine national minorities (those having a kin state: Armenian, Belarusian, Czech, German, Jewish, Lithuanian, Russian, Slovak, and Ukrainian), four ethnic minorities (i.e., those without a kin state: Karaim, Lemko, Roma, and Tatar), and the regional language (Kashubian), granting them a range of political/linguistic rights (for more, cf. Wicherkiewicz 2014).

This diversity of non-Polish languages complemented the richness of the dialects and varieties of the Polish language itself. At that time, when they were largely unexplored, they must have diverged from each other to

a much greater extent than their contemporary dialectological distance and research might suggest (scholars began to be interested in Polish dialects only at the beginning of the 19th century).

The structure of the present paper is as follows: Section 2 of this paper provides general information about the initial assumptions and main goals of Poland's Linguistic Heritage Database. In Section 3, the structure of the contents and the language inventory included in the database are described, and Section 4 is devoted to selected issues of database implementation and development. Finally, conclusions and final remarks are formulated in Section 5, together with a brief discussion of future work.

2. Design, Assumptions and Goals

Poland's Heritage Database was developed within a project called *The Linguistic Heritage of Rzeczpospolita* ('Dziedzictwo językowe Rzeczypospolitej' in Polish). As discussed above, we have deliberately referred to the hardly translatable concept of 'Rzeczpospolita' – used with reference to various stages of Poland's statehood in history. The general initial assumptions of the project were to explore and describe the linguistic diversity of the territory, which resulted mainly from the multilingual, pluriethnic and de facto multinational dimensions of that political body.

Based on these assumptions, the most important goals of the project were formulated as follows:

- to provide information about languages for a general audience (over twenty languages and/or language varieties described and included in the present version of the database);
- to start a scientific documentation database providing resources for researchers; primarily linguists and language documenters but also culture anthropologists and others (four languages featured so far: Latgalian, (Polish) Yiddish, Vilamovian (Wilamowicean), and Hałcnovian).

3. Information Structure and Language Inventory

The main criteria for choosing the given language varieties for their documentation and archiving were their co-existence on the territories of the Republic (or, perhaps more suitably, Republics) of Poland and, more importantly, the fact that they have all been affected by contact with the Polish language in modern times. A number of the dialects, languages and

varieties finally included in the presented heritage database are already extinct and the others are at various, but serious, stages of endangerment – therefore they ought to be regarded as (highly) vulnerable. The design of the database structure assumed the possibility to store both general descriptive information about languages as well as more elaborate or detailed reports and source materials. The languages for which such types of extended information have been provided are labelled as 'featured languages' in the database. At the current stage of development (2014), the database delivers information for a total of 22 languages and varieties (the complete list can be inspected in Section 3.1). Four of the languages have received the 'featured language' label (see Section 3.2.).

3.1. Language Profiles

The database provides elaborate encyclopaedic information (henceforth named a language profile) concerning the language typology, history, writing system(s), users, characteristics and geographical location of the speakers' community, cultural issues, the level of endangerment (according to the UNESCO classification[10]), and the status of existing documentation and literature for each of the following languages:

- Western Belarusian dialects,
- Belarusian-Ukrainian transition varieties of Podlachia,
- Polesian varieties – spoken and written in Poland and Belarus,
- Old-Believers' Russian,
- Western Ukrainian dialects,
- Rusyn / Lemko,
- Czech varieties spoken in enclaves of Zelów, Kuców (nonexistent) and Kłodzko Valley,
- Lakhic – transitional varieties of Polish / Czech / Moravian borderland,
- Polish-Slovak varieties of Orawa and Spisz,
- Latgalian,
- Lithuanian dialects (Dzukian, Suvalkian),
- Low German, including Mennonites' Plautdietsch, spoken in the present-day northern Polish territories until the 1950s,
- High German dialects of Silesia and Central Poland,
- Germanic language exclaves – mainly Bielsko-Biała *Sprachinsel*, including the still spoken ethnolect of Wilamowice,

[10] Mainly according to the above-mentioned UNESCO *Atlas of the World's Languages in Danger* and/or estimations by the present research team.

- Polish Yiddish,
- Romani tribolects,
- Armeno-Kypchak – a Turkic variety used in the past by Polish Armenians,
- Karaim dialects, spoken by the Polish Karaims in two varieties: Trakai and Lutsk-Halych (in Poland and Western Ukraine practically extinct; still maintained in the Republic of Lithuania), and
- Tatar (and their Polish-Belarusian regiolect written with Arabic script) – extinct.

The language profiles are illustrated with maps (the majority of which were created specifically for the needs of the present database), illustrations, photographs, and diagrams.

3.2. Featured Languages

The database makes it possible to extend the information about each of the represented languages by attaching source materials, elaborate descriptions and analysis results to the existing language profiles. So far, such extended information, accompanied by instruments for comprehensive and representative documentation and professional linguistic annotation, have been provided for four featured languages and varieties:

- Latgalian,
- Polish Yiddish,
- Wilamowicean (Vilamovian, Wymysörys), and
- Hałcnovian (as the remnants of the Bielitz-Bialaer Sprachinsel).

A brief description of the featured languages and the related newly collected resources follow below.

Latgalian – a Baltic language spoken in the region of Latgale (Eastern Latvia), known also as Polish Livonia (1620s-1772 a province of Rzeczpospolita). In the past, Latgalian developed under a durable language contact with Polish, although the ethnolinguistic ties were broken in the 19th century. Among others, the Latgalian materials in the present database include a collection of 19th-century Latgalian texts by S. Ulanowska (a Polish ethnographer and fieldwork linguist) recently annotated, re-consulted with native speakers and Latgalian philologists, and partially audio-video re-recorded (cf. Leikuma 2001; Stafecka 2003; Nau 2008; Nau 2009; Nau 2011).

Polish Yiddish – once the widest-used variety of the Yiddish language; before the Holocaust, spoken as a native language by more than 3 million Polish Jews. The language developed vivid, abundant and mutual language contacts with Polish, testified also in the written/literary language. It is not only the extermination of the Polish Yiddish language community but also language planning decisions (basing the modern Yiddish standard on Litvish / Lithuanian Yiddish) that contributed to the critical endangerment of Polish Yiddish (cf. Birnbaum 1979; Jacobs 2005; Weinreich 2006). The Polish Yiddish materials in the present database comprise audio-video samples of recorded texts (biograms, stories, memories) as well as fragments of literature in Polish Yiddish.

Wilamowice/Wymysöu and **Hałcnów/Alza** are the last two Germanic language exclaves in Poland, the remnants of the more extensive Bielsko-Biała (Bielitz-Biala) Sprachinsel established in the 12th to the 13th century during the colonization of Silesia, and constituting an archaic example of Middle High German exterriorial and diachronic varieties, developed in contact with the Polish-language environment. Wilamowicean is now spoken by fewer than 50 users, including only two representatives of younger generations, while Hałcnowian is remembered by the last seven (semi-)speakers (cf. Majewicz 1991; Wicherkiewicz and Zieniukowa 2001; Wicherkiewicz 2003; Gara 2003). The materials from Wilamowice and Hałcnów included in the present database consist of various video and audio recordings (biograms, songs, memories, myths, stories, poems, lists, dialogues, scenes, literary texts, written memoirs, and photographs).

4. Database Implementation and Development

The database was implemented with the use of SQL Server technology. In response to the need for a database edition by non-programmers (linguists, language documenters, phoneticians, students), an Internet data-management application was developed using ASP.NET. The online application supports defining multiple users and simultaneous edition and modification of the database contents.

4.1. Application Structure

Internally, the application was divided into three separate programming layers: *data*, *logic* and *presentation*. The *data* layers correspond to a relation database created in a freeware SQL Server Express, ensuring efficiency and automatically providing data security. The middle layer, *logic* (also named business layer), includes a range of object classes in C#

language representing data and used at the stage of programming. This enables fully object-oriented programming and guarantees avoiding redundancies. The *presentation* layer represents controls and websites created in ASP available directly to the end-user. This solution made it possible to develop:

- the online editor of the database (Section 4.2), available for authorized users, and
- the publicly accessible website for data presentation and sharing (Section 4.3).

Both the editor and the publicly accessible website were created using exactly the same relation database, without any need for duplicating or copying data.

4.2. Database Editing: Online Editor

The application was installed on an online server, which enabled simultaneous access to the database by multiple users. The use of the ASP.NET technology made it possible to program the website with C# programming language. The implemented solution does not require installing any additional software in order to access the data. Any Internet browser can be used to edit contents with an application comparable in ergonomics and ease of use to desktop applications (an example screenshot is shown in Figure 2).

The data contributing to the database have been divided into three main categories: (1) language profiles (as described in Section 3.1), (2) source texts, and (3) source 'phrases.' The texts are connected with their respective language profiles while phrases are connected with the texts from which they are derived. 'Phrases' are the smallest content entries in the database, corresponding to the shortest fragments of the source texts. In the process of translating the original texts, information for each phrase is stored separately with a view to enabling flexibility in displaying the data (e.g., an interactive display of texts by highlighting the original phrases along with their transcriptions, transliterations or translations). Due to the huge variability of character encodings used in the represented languages, a Unicode font was used in all editing forms of the database (Lucida Sans Unicode).

Another option implemented in the data-management application is the possibility to attach multimedia files (audio, video, pictures, PDF files, etc.) attributed to any of the three categories of data; that is, the attachments can become related to the selected language profile or, more

specifically, to a source text or even to a single component phrase. All multimedia files are stored in their original formats. Even if they were temporarily converted to other formats or divided into smaller parts, the original versions are kept in the repository (cf. also Gibbon et al. 2004). The multimedia files collected for the featured languages (Section 3.2) were prepared (annotated, subtitled) with Elan (Sloetjes and Wittenburg 2008) (audio-video recordings) and Annotation Pro (Klessa et al. 2013) (audio recordings). Both software tools produce XML-based annotation data formats.

For each type of data (language profiles sections, source texts, source phrases and their respective attachments), it is possible to add metadata information (related to the features of particular content items, their authors and contributors, technical information, etc.). Metadata are stored in separate, searchable fields.

Figure 2. On-line database editor available for authorized users – user interface (example view).[11]

4.3. Data Presentation and Sharing

The final product of the project is the *Poland's Linguistic Heritage Database* website developed for data presentation and sharing (also implemented with ASP.NET) of all data contributed to the database via the online editor.

[11] Figure 2 shows an example screenshot from the on-line database editor designed specifically for the present project, available on-line for authorized users at http://inne-jezyki.amu.edu.pl/editor/

Data can be accessed at the address: *www.inne-jezyki.amu.edu.pl* (an example screenshot is shown in Figure 3). The English equivalent for the name "inne-jezyki" used in the website address might be "different languages" or "other languages." As mentioned above, the territory of the Republic of Poland has always been inhabited by communities speaking languages different from Polish, particularly during the period of "Rzeczpospolita," hence the website name.

The main assumptions for creating the website were to achieve data transparency, thus making the contents available for the general public, and at the same time preserve the possibility to modify and further extend the contents of the database. The implemented functionality enables an advanced search through the data and flexible, interactive visualization of the contents.

Two language versions of the final website have been implemented; that is, Polish and English. All language profiles are available in both languages. Additionally, a part of the newly collected source materials for the featured languages was translated into Polish and English (c.f. for example, selected texts in Yiddish or the texts originally collected by S. Ulanowska now translated from Latgalian to Polish and English).

4.3.1. Accessibility and levels of permissions

One of the basic assumptions was that all language profiles ought to be freely available for all users while the availability of the source materials and research results was regarded as dependant on permissions granted by the authors of particular items or other copyright holders. With a view to modelling these authorization requirements, three levels of permissions were implemented in the database:

- Level 1. Public – available for all users
- Level 2. Partly restricted – available for all registered users, e.g., for educational use
- Level 3. Private – only for selected, individually authorized persons