A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages

Radovan Garabík[1], Markéta Caravolas[2], Brett Kessler[3], Eva Höflerová[4], Jackie Masterson[5], Marína Mikulajová[6], Marcin Szczerbiński[7], and Piotr Wierzchoń[8]

[1] Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
[2] School of Psychology, University of Liverpool
[3] Psychology Department, Washington University in St. Louis
[4] Department of Czech Language and Literature with Didactics, University of Ostrava
[5] Institute of Education, University of London
[6] Faculty of Education, Comenius University, Bratislava
[7] Department of Human Communication Sciences, The University of Sheffield
[8] Institute of Linguistics, Adam Mickiewicz University, Poznań

Abstract. We describe a lexical database consisting of morphologically and phonetically tagged words that occur in the texts primarily used for language arts instruction in the Czech Republic, Poland and Slovakia in the initial period of primary education (up to grade 4 or 5). The database aims to parallel the contents and usage of the British English Children's Printed Word Database. It contains words from texts of the most widely used Czech, Polish and Slovak textbooks. The corpus is accessible via a simple WWW interface, allowing regular expression searches and boolean expression across word forms, lemmas, morphology tags and phonemic transcription, and providing useful statistics on the textwords included. We anticipate extensive usage of the database as a reference in the development of psychodiagnostic batteries for literacy impairments in the three languages, as well as for the creation of experimental materials in psycholinguistic research.

1 Motivations for the West Slavic database

Lexical databases that reflect language use across the developmental spectrum are critical tools for research on the development of spoken and written language skills because they allow researchers to select materials for their studies that are age- and grade-appropriate. A number of databases exist for adult language, but only a few have been developed based on child language. Available child-language corpora include the earlier American English sources *The American Heritage Word Frequency Book* (Carroll, 1971) and *The Educator's Word Frequency Guide* (Zeno, Ivenz, Millard & Duvvuri, 1995), and more recently, *Manulex*, a French database (Lété, Sprenger-Charolles & Colé, 2004) and the British English *Children's Printed Word Database* – CPWD (Masterson, Stuart, Dixon & Lovejoy, 2003). However, to our knowledge, no corpus of children's printed words has been published in any of the Slavic languages.

The data that can be generated from lexical databases have diverse applications in psycholinguistic research. For example they can produce statistics about lexical and sublexical variables such as frequency of specific units, word length in terms of letters or syllables, orthographic and phonological neighbourhoods, and grapheme–phoneme consistency, to name but a few. Accumulating evidence shows that text-based variables such as these affect learning to read and spell from an early age (e.g. Caravolas, Kessler, Hulme & Snowling, 2005; Treiman & Kessler, 2006; Pacton, Perruchet & Fayol, 2001). An emerging key issue in this research area concerns the relative influence of orthographic depth on the learning process: Does the predictability (transparency) of a specific writing system significantly influence the way children learn to read and write it? Direct cross-language comparisons based on corpus statistics will play a critical role in answering this question. However, a current limitation is that there are still few children's lexical databases in different languages, and those that do exist, rarely generate directly comparable statistics. This is because databases may be designed for different scientific purposes and thus do not always contain similar information from language to language. Moreover, linguistic features that are important in one language may be deemed to be of marginal importance and thus not warrant inclusion in another. Thus, a fundamental motivation for our project was to redress these shortcomings in the creation of a database that would allow direct cross-linguistic comparisons of a wide range of measures across Czech, Polish and Slovak. Cognizant of the prevalence of English-language research and of English-based models of language and literacy development, we based the West Slavic lexical database (Weslalex) on the existing English CPWD (Masterson et al., 2003). These design features will enable researchers of Slavic languages to investigate questions that could not be addressed without corpus data, and, they will facilitate meaningful comparisons to English measures, which so often provide the benchmark in developmental psycholinguistic research.

2 Types of corpus statistics provided in existing children's corpora

The existing American English corpora provide only word frequency information across (Carroll, 1971) and within (Zeno et al., 1995) primary school grades. The French Manulex (Lété et al., 2004) currently contains lemmatized and nonlemmatized grade-level word frequency lists, limited part of speech (POS) information, and letter frequencies. The more recent extension, Manulex-infra (Peereman, Lété & Sprenger-Charolles, in press), generates statistics at the sublexical level (syllable, grapheme-to-phoneme mappings, bigrams), and lexical level (lexical neighborhood, homophony and homography). The British English CPWD (Masterson et al., 2003) allows searches by grade and it offers a wide range of possible searches at the lexical and sublexical levels. These include searches of orthographic and phonological attributes such as neighbourhoods, component letters and phonemes, word length, and frequency. A feature that is currently missing from all of these corpora is a detailed morphosyntactic level of analysis.

Although impressive advances are now being made in several languages, no comprehensive children's database, that includes all of the above search possibilities, has yet been developed.

3 Features of the West Slavic database

The database that we are developing is modelled in part on the CPWD, and one of our key objectives is to make possible parallel cross-linguistic searches in any of our Slavic languages and this English language resource. In addition, however, a truly useful tool for psycholinguistic research in the inflected Slavic languages requires information not only at the lexical and sublexical (grapho-phonological) levels, but also at the morphophonological, grammatical and phrase levels.

Thus we include POS information derived from sentence-level analyses, and one of our search tools permits searching of multiword sequences. The integration of the lexical/sublexical database and of the sentence-level database is one of the critical challenges being addressed in our project.

4 A description of the pilot database materials

The database currently contains printed words in Czech (388 654 tokens, 64 411 distinct wordforms, 24 364 distinct lemmas), Polish (175 404 tokens, 34 067 distinct wordforms, 13 767 distinct lemmas), and Slovak (180 674 tokens, 30 060 distinct wordforms, 14 610 distinct lemmas)[1] from texts primarily used for language arts instruction in each country in grades up to 4 or 5. Based on surveys carried out in each of the three countries, we selected books and materials from those series that are currently the most widely used. Some intercultural differences necessarily emerged so that different numbers of books were sampled in each country. The simplest case proved to be Slovak where only one language arts series is approved by the Ministry of Education; thus we selected the designated readers and one Slovak language grammar book from each primary grade (1 to 4). The total number of Slovak books is therefore relatively small (9 books), but they represent an exhaustive sample of the materials children read as part of their language arts instruction. The Czech case was less straightforward because several Ministry-approved series exist; however, we chose the two series that predominate. Thus for each grade level (1 to 5) we chose one reader and one grammar workbook from each series for a total of 19 books (grade 1 did not have a grammar workbook). The Polish case was the most complex for two reasons. First, as in Czech, several series are Ministry-approved, thus necessitating the selection of a sub-sample. Second, the recently reformed primary education curriculum integrates teaching of different subject areas and thus no separate, dedicated language arts text books are currently in use. Instead, for

[1] The count contains the tokens without punctuation, digits and nonwords. Also the text annotated as instructions is excluded from the lowest grade – the reason being that these instructions are presumably not read directly by the children in this grade.

each grade level (0–3 equivalent[2] to 1–4 in the Czech and Slovak school system) children receive up to 20 booklets in which language as well as maths, science, etc. are covered in overlapping sequences. Consequently, we selected one widely used scheme and within this we selected 11 booklets (five from grade 0, two from grades 1–3 each), we prioritized those with a greater emphasis on language arts and reading where possible. We are confident that using these procedures we have sampled books in each language that are highly representative of the reading materials encountered by primary school children.

5 Text processing and annotation

All of the books were scanned and submitted to OCR. The OCR-ed text files were then proofread and annotated by proficient speakers of each language (typically students in language education or in psycholinguistics). The texts were then analysed for morphological categories and POS in each language, and were phonologically transcribed, obtaining texts with POS, morphological categories and phonemic transcription for each word.

The texts were manually annotated with XML to mark nonwords and metatext – instructions on using the text. Using XML tags provides the possibility of using already existing tools for XML validation, thus reducing the number of annotation mistakes.

6 Morphosyntactic annotation

Due to the highly inflected character and rich morphology of the languages in question, morphosyntactic analysis and lemmatization is not a trivial task. Given the number of texts present, it was impractical to annotate the words manually, and we had to use automatic tools. For Czech, the tagger described in (Hajič & Hladká, 1997) was used.

Since no sufficiently accurate Polish parser was freely available, we used the Waspell (Płotnicki, 2003) morphological analyzer program, which is based on the Ispell Polish dictionary.[3]

For Slovak, although another morphological tagset and tools for automatic text processing (Garabík, 2006) exists, we decided to use the same tagset as the Czech one, adapted for the Slovak language (Hajič, Hric & Kuboň, 2000). This will make eventual cross-linguistic comparison between Czech and Slovak texts easier.

6.1 Czech and Slovak tagsets

The Czech tagset describes 13 different morphological categories: part of speech, detailed part of speech,[4] gender, number, case, posessor's gender, posessor's number, person, tense, degree of comparison, negativeness, voice and register. There

[2] Children at the age of 6 enter a reception grade, which is referred to as grade 0.

[3] `http://ispell-pl.sourceforge.net/`

[4] As named by the authors.

are 12 different main part of speech categories, with particles forming their own category (as is the custom in Slavic linguistics). Punctuation also has its own tag (including sentence boundaries mark).

Special care has to be taken in interpreting the gender category – in addition to four common values (masculine animate, masculine inanimate, feminine and neuter), there are also additional possible values, corresponding to genders conflated due to inflectional syncretism. These are "feminine or neuter", "feminine singular only or neuter plural only", "masculine inanimate or feminine plural only", "masculine (either animate or inanimate)", and "not feminine".

The number category contains a special value for the old Slavic dual, present in the Czech language only in a few nouns (not present in the Slovak version of the tagset). The most notable deficiency of the tagset is the absence of a verb aspect category.

6.2 Polish tagset

The Waspell morphological analyzer analyses wordforms in isolation, without taking syntactic context into account. While it produces a detailed morphological description (similar to that obtained for the Czech and Slovak corpora) it results in a large proportion of alternative descriptions, due to inflectional syncretism. For example, each occurrence of the token *miał*, which can mean either 'coal dust' or 'he had', receives these three analyses:

- noun, masculine, singular, nominative
- noun, masculine, singular, accusative
- verb, perfective, indicative, past tense, 3^{rd} person masculine singular

Manual disambiguation of this output (by identifying ambiguous words in original texts) was not feasible at the level of full morphological description. Therefore, for the pilot version of the database we decided to specify only the part of speech. While part of speech could also be ambiguous (as in the example given above) this ambiguity affected only a small proportion (1–2%) of wordforms. Disambiguation was accomplished by listing all ambiguous wordforms in Waspell's output, and checking them against the original text.

7 Phonemic transcription

It was desirable to include a phonemic transcription of the written texts. This gives us access to various statistical analyses on the spoken, rather than the written, language level. There were, however, several open problems concerning the exact nature of the transcription. One possibility was to deploy the SAMPA transcription (Fourcin, Harland, Barry & Hazan, 1989), a substitute for the International Phonetic Alphabet which has the advantage of using only ASCII characters, which are easily entered and do not require any special software arrangements at the client's side, such as special fonts and keyboard layout. However, with modern computing systems, the technical advantages of SAMPA diminish and some disadvantages come to the fore. The SAMPA transcription is usually specific for a given language, and transcriptions for different languages sometimes collide, complicating further comparisons (however, this is not the case with the Czech, Polish and Slovak transcriptions). Moreover, SAMPA for these languages uses a rich variety of non-alphanumeric symbols, interfering with regular expressions, thus complicating queries in the database. We therefore opted for an IPA transcription, using Unicode internally and presenting the output in UTF-8 encoding.

Pronunciation was rendered in terms of classical phonemic analysis, which means that sometimes salient phonetic and morphological information was disregarded. Diphthongs were treated as a sequence of two phonemes (vowel plus glide), affricates and long vowels were treated as single phonemes. When no confusion could result, typographically simple (preferably ASCII) IPA characters were chosen over more complex and precise ones: thus Czech *hezká* 'pretty' is transcribed as /heskaː/, not /ɦɛskaː/. Nonsyllabic components of diphthongs were given distinctive representations as glides (/j/ as in *kraj* /kraj/ 'region', or /u̯/ as in Czech and Slovak *auto* /au̯to/ 'car' – disyllabic sequences of vowel plus vowel are possible in the languages targeted.

Syllable boundaries were not marked, because they are usually inferrable from the phoneme sequence; at other times they can be controversial. Word stress was not transcribed because it is not phonemic.

There are different possibilities in representing sibilant affricates. The first and simplest possibility is to encode them as the sequence of constituent phonemes, /ts/, /dz/, /tʃ/, /dʒ/, /tɕ/, /dʑ/. This is clearly unacceptable for Slavic languages, because the affricates are phonologically different from diphonemic, non-affricate sequences of phonemes. Another possibility is the obsolete IPA ligature notation: /ʦ/, /ʣ/, /ʧ/, /ʤ/, /ʨ/, /ʥ/. This has the clear advantage of representing one (affricate) phoneme with one Unicode character, facilitating regular expression queries and statistical analysis. The last possibility to be considered is to use the official IPA transcription (with the tie bar above), /t͡s/, /d͡z/, /t͡ʃ/, /d͡ʒ/, /t͡ɕ/, /d͡ʑ/. This has the advantage of being official, but also two principal disadvantages. The first is that there is not yet quite uniform support of Unicode combining characters across various operating systems and font rendering engines commonly used. However, support is rapidly growing and is present in all the recent common computing environments, and, where lacking, an easy, free upgrade is almost always available. The second disadvantage is the fact that each affricate is now represented as a sequence of three Unicode characters. This makes the queries more difficult, e.g. to search for a word with a phoneme /t/, instead of writing a regular expression `.*t.*`, we have to make sure that the character following the character after /t/ is not a `U+0361 COMBINING DOUBLE INVERTED BREVE`. The above mentioned regular expression would look like `.*t(?!.͡).*` – that is, /t/ not followed by two characters, the second of which is the tie bar, followed by any sequence of characters. This is complicated by difficulties of entering and editing text with a standalone combining character; however, these complications can be remedied by putting a special translation level in the WWW interface, translating simpler (preferably noncombining), user entered characters into complicated IPA Unicode character sequences.

7.1 Czech phonemic transcription

A standard literary Czech pronunciation was chosen as a reference. In the absence of definitive, freely available pronunciators, Czech pronunciations were algorithmically inferred from their spelling. Proceeding from left to right through the word, each of approximately 94 context-sensitive rules were consulted, and the first rule that fit the context determined the pronunciation of the letter in question. For example, the rules for the letter *ě* decode it as follows:

1. /ɲe/ if immediately preceded by *m*
2. /e/ if immediately preceded by a dental stop letter, i.e., the regular expression `[dnt]`
3. /je/ otherwise

In turn, the rules for the dental stop letters each contain a clause mapping them to a palatal stop when followed by *ě*, among other letters (i.e., `[ěií]`). The contextual rules all reference the input orthography, not the output phonology, and there are no multi-step derivations for individual letters. A few rules make special provisions for consuming more than one letter at a time (*ch* mapping to /x/, *dz* mapping to /d͡z/, *dž* mapping to /d͡ʒ/, but such complications are rare in Czech, which has very few digraphs.

Most of the context-sensitive rules serve to handle voicing issues, such as when letters like *d*, normally voiced (/d/), are devoiced, as in *led* /let/ 'ice', or vice versa *četba* /t͡ʃedba/ 'reading'. Such mappings can easily be predicted from the immediate context. A slightly more complicated case is that devoicing occurs before final *me* but only in first-person plural imperative verb forms, as in *odpovězme* /otpovjesme/ 'let us answer'. Here the rule system can exploit the fact that the words entering the pronunciation module have already been tagged for morphosyntactics; analysis tags beginning `Vi-P---1` identify words as being first-person plural imperative verbs.

Unfortunately, however, the tagging system does not provide all the information needed to unambiguously apply all of the rules of Czech orthography. For example, in the word *odzátkovat* 'to uncork', the letters *dz* are to be interpreted as two individual phonemes, /dz/, not as the digraph pronounced /d͡z/. This could be inferred if the system knew the word has a transparent morpheme boundary between the *d* and the *z* (*od-* is a private prefix, *zátka* is 'cork'), but that information is not supplied. Furthermore, there are hundreds of words whose correct pronunciation could easily be inferred if the system only knew that they were of non-Slavic origin. For example, *n* is normally pronounced /ɲ/ before *i*, but not in Latinate words like *penicilin* /penit͡silin/ 'penicillin'. These and other issues are handled with an exception list. Virtually all of the exceptions are of the form `penicilin` → `p=p e=e n=n`, which means that if a word has been tagged as a form of the lexeme *penicilin*, then if the wordform in question begins *pen*, those three letters should be assigned the phonemes /p/, /e/, and /n/, respectively, and the rest of the word should be decoded using the ordinary rules of Czech orthography.

Czech is sufficiently regular that this simple system is basically adequate, although it would be more satisfying if such "exceptional" pronunciations could be worked out from first principles (e.g., etymological information) instead of by stipulation. Catching exceptional pronunciations is fairly labour intensive, although the work was speeded significantly by checking a sorted list of the word types in the corpus against two Czech language works that include notes on some exceptional pronunciations: *Pravidla českého pravopisu* (2005) and *ABZ slovník cizích slov* (2006).

Further progress in correcting exceptional spellings will require careful attention from linguistically trained native speakers. Especially with lesser-known or more recently introduced foreign words, it is not always easy to guess how their pronunciation will be adapted to Czech phonology. Another continuing problem involves the pronunciation of unusual clusters such as double letters. Whether these are reduced to simpler pronunciations, such as single phonemes in the case of double letters, depends on a combination of factors such as the salience of any intervening morpheme boundary and a rather nebulous perception of which of multiple variant pronunciations is currently considered too colloquial or too stilted to be considered standard. In the absence of authoritative reference books, the active assistance of a skilled orthoepist is required.

Pronunciations are internally stored in a way that explicitly aligns them to their spelling. For example, the pronunciation of *hezká* 'pretty' is stored in the form `h=h e=e z=s k=k á=aː`, with letters to the left of each equals sign and phoneme representations to the right. This representation facilitates the compilation of statistics on spelling consistency and enables the searcher to easily request, for example, all words where /z/ is spelt *s*. This notation accommodates the relatively rare instances where a sound is spelt with two letters, as in *Čech* 'Czech man' `č=t͡ʃ e=e ch=x` (technically *ch* is considered a single letter in Czech) and *Anglie* 'England' `a=a n=n g=g l=l i=ij e=e`.

7.2 Polish phonemic transcription

The Polish pronunciator (which is still under development) is modeled very closely on the Czech pronunciator described above: it algorithmically derives pronunciation of single words from their spelling.

Building the pronunciator required choosing one model of "standard Polish pronunciation" among several available alternatives. Some particularly important phenomena include the varying pronunciation of word-final ę (pronounced as nasal vowel /ẽ/ in careful, conservative speech, but typically is oral /e/); the distinction between dental nasal [n] and velar nasal [ŋ], which is phonemic in the Warsaw dialect of Polish, but phonetic in the Poznań-Kraków dialect, where [ŋ] is only a positional variant of /n/ (Strutyński, 1997); and the distinction between some palatalized consonants (e.g. [p^j], [k^j]), and their non-palatalized equivalents ([p], [k], etc.) which is again either phonemic or merely phonetic, depending on the region.

Since Polish has several digraphs, (e.g. *rz*) where Czech and Slovak have accented letters (e.g. *ř*), this opens up more possibilities for parsing ambiguities; for example, *rz* represents two separate sounds in *marznąć* /marznoɲt͡ɕ/ 'to freeze', rather than its usual /ʒ/. In practice, however, we have found that very few exceptions arise if the program always treats digraphs as such; such exceptions can be handled by explicitly listing exceptional lemmas. As was the case in Czech and Slovak, the identification of such exceptional pronunciations requires checking the output of the analyzer by linguistically trained native speakers.

7.3 Slovak phonemic transcription

For the automatic Slovak transcription, we used the system described by (Ivanecký, 2003). The software transcribes the text into SAMPA (Ivanecký & Nábělková, 2002); we created translation tables into IPA as used by our database. The software can take into account consonant assimilation across word boundaries; however, we decided not to use this possibility for the pilot version, since at least initially, most queries will concern only isolated words, and because we will thus maintain compatibility with Czech and Polish transcriptions. However, devoicing at the ends of words was still applied. One problem concerns the transcription of the vowel *ä* and the syllables *le*, *li* and *lí* – all of which have two possible pronunciations, /æ/ vs. /e/; /ʎe/, /ʎi/ and /ʎiː/ vs. /le/, /li/ and /liː/. While the former pronunciations have for all practical purposes disappeared from standard Slovak, they are still considered formally correct, and as such are taught in elementary schools. Thus, using /e/ and /l/ for the transcription would accurately reflect standard spoken language, while using /æ/ and /ʎ/ would reflect the prescribed school usage. We have chosen the official prescribed variant, since teachers are more likely to pronounce the phonemes in the "correct" way, especially when teaching the letter–sound correspondences.

The accuracy of the transcribed phonetic assimilation heavily depends on the detection of intra-word morpheme boundaries, which in turn requires a good morphological dictionary, which is not currently available. Another problem is a (relatively) huge number of orthographic exceptions in Slovak, especially in loanwords. These mostly involve *e*, *i* and *í*, which routinely mark preceding *t*, *d* or *n* as palatal in native words but not in non-Slavic loans. Currently, the software contains just a small dictionary of orthographic exceptions, and therefore transcription of most of the loanwords marks palatal consonants incorrectly.

8 Query interface

We made available two independent query mechanisms, accessible from the project's page at `http://spell.psychology.wustl.edu/weslalex`. One indexes the data and presents the query results in a keyword-in-context (KWIC) interface; the other is similar to the CPWD functionality.

For the first tool, the files were converted into so-called vertical files, where each word in the text is represented by a separate line. Each line contains several tab-separated fields: the word's spelling, lemma, morphological tag, and pronunciation in IPA. This format is easily parseable by simple computer programs and therefore is well suited for custom statistical analyses beyond the capabilities of the systems described here.

The vertical files were indexed using the Manatee corpus manager system (Rychlý, 2000). It is possible to query the corpus using specialized multiplatform client software *Bonito*, offering a rich set of query possibilities and statistical analysis. It is possible to query individual token attributes matching given regular expressions, to search for sequences of arbitrary tokens, to apply negative and positive filters depending on context to search results, to sort the concordances by different criteria, and to obtain a frequency analysis of any token: raw frequency, collocates, and measures of mutual independence (raw MI scores and frequency-adjusted /t/ scores). A simplified user interface has been built to accommodate the need for quick access to queries, without the need for complex statistical analysis. This interface is accessible through a simple WWW interface, and affords the possibility of the same type of queries as Bonito. This interface contains a simple on-screen virtual IPA keyboard, to facilitate queries in the phonemic transcription.

The other query tool is intended to provide for the Czech corpus a functionality similar to that of the CPWD. It is an HTTP-based service that lets users query the corpus by desired lexical characteristics. Users can specify which specific texts they wish to search, whether they wish to include metatextual instructions, what lexical properties they wish to search by, and what properties they wish to see displayed for each retrieved word type. In addition, token counts are always provided for each word type, as well as certain types of statistics for each displayed property.

The basic properties defined for each word are its case-sensitive spelling; its lemma, lemma properties, and morphosyntactic analysis, as supplied by the (Hajič & Hladká, 1997) program; and its letter–phoneme alignment. For convenience, several secondary properties are derived when users refer to them. These include an uppercase version of the spelling, which is useful when users wish to ignore case distinctions; the pronunciation; various fields for different parts of the morphosyntactic analysis, such as part of speech, gender, and so forth; and spelling and pronunciation lengths.

An unusual characteristic of the retrieval tool is that the definition of a word type is not built in, but is defined by the end users based on what properties they ask to have displayed. For example, if users request case-sensitive spelling, lemma information, and full morphosyntactic analysis, the definition of word type will be very specific. A word like *růže* 'rose(s)' could appear as several different word types depending on the specific case and number it represents (nominative singular, genitive singular, nominative plural, etc.), and a capitalized version would represent many more types. On the other hand, if users request only the uppercased version and do not ask for full morphosyntactic analysis, all uses of *růže* or *Růže* will be subsumed under only one word type.

The system also displays aggregated statistical information depending on what fields the users ask for. If users ask to see a spelling field, the tool will give the range and mean of the word lengths and will also tell how often each letter appeared. If the letter-phoneme alignment field is requested, spelling correspondence statistics are provided in each direction: e.g., how many times the letter *z* represented each of its possible pronunciations (/z/ and /s/), and how many times a phoneme, such as /z/, is spelled with each of its possible spellings (*z* and *s*). For the categorical fields, such as gender, a tally is provided telling how many times each individual level, such as feminine or masculine, appeared. All counts are given both by word types and by word tokens.

By default, information is presented about all the words in the corpus. It is also possible to limit the search to words that have particular properties. Typical boolean search queries are accepted. Most fields contain character strings and can be searched via arbitrary regular expressions. For example, the following query looks for nouns that end in *-o* in the nominative or accusative case but which are (contrary to the usual pattern) not neuter: `spell_uc = ".+O" and pos = "N" and (case = "1" or case = "4") and not gender = "N"`

Output is sorted in Czech lexicographic order and is presented as a list of tab-separated values, to facilitate importation into statistical programs or spreadsheets. The default character encoding is Unicode UTF-8, with spellings presented in normal Czech orthography. However, because it is not always convenient for users to input such characters, ASCII sequences are also understood, such as *c<* for *č* and *u0* for *ů*. For IPA characters, several different synonyms are often accepted: /t͡ʃ/ can be input as /t_s</ or /c</, or, for those with Czech keyboards, /t_š/ or /č/. Display uses the full Unicode character set unless the user specifically requests a more limited set; choices are ASCII and two character sets often used for Central European computing: Windows-1250 and ISO/IEC 8859-2. These are provided primarily to accommodate legacy software.

Although the tool is currently very usable, several enhancements are being planned, foremost among these being the inclusion of material from the Polish and Slovak corpora. We also plan several additions to bring the interface and capabilities more in line with those of the CPWD. This involves adding new derived properties, such as counts of orthographic neighbours (words that differ by only one letter). Perhaps even more important will be the addition of simplified search forms to make the corpus easy to search by people who have no prior experience with boolean search techniques and who have not read the help files.

9 Applications

The database was conceived by developmental psycholinguists and linguists for primary use in various types of developmental research. One important application of children's printed word statistics is in the study of the impact of 'exposure to print' and of implicit learning mechanisms on the development of reading and spelling skills (e.g., Cassar & Treiman, 1997; Steffler, 2004). For example, in research on spelling development in a language like Slovak, it is important to understand the effectiveness of explicit instruction on children's learning of rules such as that governing the spelling of palatalized consonants (e.g., /c/ is spelt *t* before *i*, *í*, *e* but as *ť* in other environments), and to what extent this learning is influenced simply by the frequency of exposure to constructions that reflect this rule; there are many more words with *t*, such as *teta* 'aunt' than with *ť*, such as *ťava* 'camel'. Until now, only adult-language corpora were available to address this question. However, it is not always appropriate to rely on adult-language statistics. To illustrate, for a high frequency word in child language such as *teta* (in Slovak /ceta/), which is a model word used to teach the spelling rule for the palatalized consonant /c/ in second grade, the adult frequency according to the Slovak National Corpus is 23 occurrences per million (according to Mistrík, 1969 it is 3 occurrences per million). In contrast, the real frequency of *teta* in children's printed materials is much higher — in our Slavic database, it is 175 occurrences per million. Moreover, this word is acquired early, around 2.6 years of age (based on age of acquisition norms derived from adult Slovak speakers' ratings (Mikulajová, in preparation)), and thus is highly familiar to children by the time they are in second grade. Consequently, children might learn to spell /ceta/ correctly as *teta* and not *ťeta* through exposure to print during activities like shared reading of fairy tales and bedtime stories, well before the rule is explicitly taught.

The database is already being put to use in a broader project aimed at the creation of the first diagnostic spelling test in the Slovak language (Caravolas, Mikulajová & Vencelová, in preparation). Word lists for this standardized test battery were selected according to several criteria, one of which was word frequency. The database allowed us to check the frequency of words and sublexical units that appear in the reading materials that the majority of Slovak children are exposed to in grades 1–4. Thus, the role of explicit (via classroom instruction of rules and patterns) and implicit learning of spelling rules (from exposure to print) could be compared.

In another study (Caravolas, Mikulajová & Vencelová, 2007) the database was used to estimate frequencies of different types of Slovak graphemes (with and without diacritics) and syllables to investigate the role of sound–letter consistency and letter–sound complexity in learning canonical and contextually conditioned letter spellings in Slovak.

We envisage many further uses of our database, not only for a wide variety of single-language and cross-linguistic investigations of literacy development, but also in studies of child language development, bilingualism and biliteracy (at least in the languages included in the West Slavic database and English). Importantly, we expect the database to have applications beyond the field of psycholinguistic research. For example, teachers will be able to search for appropriate word lists for teaching and remediation, and writers of children's materials will have easy access to the vocabulary that children encounter in schoolbook reading at different ages. Not least, the database can be used for advancing the work of computational linguists currently working on Czech, Polish and Slovak adult corpora, none of which currently contain sublexical or phonological information. Thus, we see our contribution of a child-language corpus as being of potential general interest to educators, linguists and psycholinguists alike.

Acknowledgement

The present project was supported by grant SG-40461 from the British Academy awarded to Markéta Caravolas, Jackie Masterson and Marcin Szczerbiński. The authors gratefully acknowledge the assistance of Vladimír Petkevič and Pavel Machač in POS tagging and phonological analysis of the Czech corpus, Małgorzata Kurek and Mária Podhradská for proofreading and mark-up of the Polish and Slovak texts.

References

ABZ slovník cizích slov (2006). Retrieved from `http://slovnik-cizich-slov.abz.cz/web.php/o-slovniku`.

Caravolas, M., Kessler, B., Hulme, C., & Snowling, M. (2005). Effects of orthographic consistency, word frequency, and letter knowledge on children's vowel spelling development. *Journal of Experimental Child Psychology*, *92*, 307–321.

Caravolas, M., Mikulajová, M., & Vencelová, L. (2007). *Effects of sound-letter consistency and letter-spelling complexity in learning canonical and contextually conditioned letter spellings in Slovak.* Poster presented at Society for the Scientific Study of Reading. Prague, Czech Republic.

Caravolas, M., Mikulajová, M., & Vencelová, L. (In preparation). Súbor testov na hodnotenie pravopisných schopností.

Carroll, J. B. (1971). *The American Heritage Word Frequency Book.* Houghton Mifflin.

Cassar, M. & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letter in words. *Journal of Educational Psychology*, *89*(4), 631–644.

Fourcin, A. J., Harland, G., Barry, W., & Hazan, V. (1989). *Speech input and output assessment: multilingual methods and standards.* New York, NY, USA: Halsted Press.

Garabík, R. (2006). Slovak morphology analyzer based on Levenshtein edit operations. In *Proceedings of the WIKT'06 conference*, (pp. 2–5).

Hajič, J., Hric, J., & Kuboň, V. (2000). Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*, (pp. 7–12)., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hajič, J. & Hladká, B. (1997). Probabilistic and rule-based tagger of an inflective language: a comparison. In *Proceedings of the fifth conference on Applied natural language processing*, (pp. 111–118)., San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ivanecký, J. (2003). *Automatická transkripcia a segmentácia reči.* PhD thesis, Faculty of Electrical Engineering and Informatics, Technical University, Košice.

Ivanecký, J. & Nábělková, M. (2002). Fonetická transkripcia SAMPA a slovenčina. *Jazykovedný časopis*, *53*(2), 81–95.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156–166.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2003). Children's Printed Word Database. ESRC research project.

Mikulajová, M. (In preparation). Age of Acquisition norms derived from adult Slovak speakers' ratings.

Mistrík, J. (1969). *Frekvencia slov v slovenčine.* Vydavateľstvo slovenskej akadémie vied.

Pacton, S., Perruchet, P., & Fayol, M. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401–426.

Peereman, R., Lété, B., & Sprenger-Charolles, L. (in press). Manulex-Infra: Distributional characteristics of grapheme-phoneme mappings, infra-lexical and lexical units in child-directed written material. Behavior Research Methods.

Pravidla českého pravopisu (2005). Academia. Kolektiv Ústavu pro jazyk český.

Płotnicki, Z. (2003). Słownik morfologiczny języka polskiego na licencji LGPL. Master's thesis, Poznań University of Technology.

Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace.* PhD thesis, Faculty of Informatics, Masaryk University, Brno.

Steffler, D. (2004). An investigation of grade 5 children's knowledge of the doubling rule in spelling. *Journal of Research in Reading*, *27*, 248–264.

Strutyński, J. (1997). *Gramatyka polska.* Kraków: Wydawnictwo Tomasz Strutyński.

Treiman, R. & Kessler, B. (2006). Spelling as statistical learning: Using consonantal context to spell vowels. *Journal of Educational Psychology*, *98*, 642–652.

Zeno, S. M., Ivenz, S. H., Millard, R. T., & Duvvuri, R. (1995). *The Educator's Word Frequency Guide.* Brewster, NY, USA: Touchstone Applied Science Associates.

Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

Computer Treatment of Slavic and East European Languages

Fourth International Seminar
Bratislava, Slovakia, 25–27 October 2007
Proceedings

Editors
Jana Levická
Radovan Garabík

Tribun

Bratislava 2007

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2007
`http://korpus.juls.savba.sk/~slovko/`