A Guide to Using the Excel Versions of the Weslalex Word Lists

Brett Kessler

2009-10-08

# Prerequisites

## Excel 2007

The word lists were designed to be opened with Microsoft Excel 2007 in Microsoft Vista. I haven’t tested them in other environments. I would strongly encourage users of earlier versions of Excel to upgrade to Excel 2007. If this is not possible, there are two limited workarounds.

1. Microsoft offers a Compatibility Pack which may allow you to open and view the files if you have an earlier version of Excel. However, the Weslalex files use several new Excel 2007 features which presumably will not work with the compatibility pack. Further, the user interface will be quite different, making many of the descriptions in this document useless.
2. This Web site provides CSV versions of the Excel files. These .csv files are plaintext: they have all the same data, but not the same formatting, usage instructions, or tie-in with Excel software. If you have your own software, or a version of Excel that cannot open the .xlsx files, you probably will be able to process these CSV files. Again, many of the descriptions and instructions in this document will not be applicable.

## Processing Extended Characters

The word lists contain special characters used in the standard orthography of Czech, Polish, and Slovak: characters such as <á>, <ů>, and <š>. In addition, the pronunciation fields contain characters of the International Phonetic Alphabet (IPA), such as <ʃ> and <ɲ>. As with any character that is not normally used for English, there is the possibility that some computer systems will not *display* them correctly. The other issue is whether you will be able to type in such characters: the matter of *input*. These two issues are treated in the two subsections below.

First, though, it is important to understand that the Weslalex files use a character encoding called [Unicode](http://www.unicode.org/standard/WhatIsUnicode.html). Character encodings assign a unique code to each possible character (letter, punctuation mark, etc.). Unicode is not the same encoding as many people are familiar with in Central Europe; crucially, it differs from such encodings as Macintosh Central European or Windows-1250 or ISO Latin-2. But it has many advantages, including the possibility of combining virtually all known languages in the same plaintext document, and the fact that it is overwhelmingly the prevailing modern international standard.

Table 1 lists all the characters used in Weslalex, along with their Unicodes. The initial “U+” on all of these codes is there simply to remind you that it is a Unicode; in many contexts all that is used is the following four numbers or letters. The characters are listed in Unicode order. Various ranges, or subsets, of Unicode have names. The Basic Latin range is essentially the same as ASCII, and covers the letters found in English as well as all the punctuation marks used in Weslalex. Latin-1 Supplement includes accented letters used throughout Europe. Latin Extended-A covers the other accented letters used only in central Europe. IPA Extensions, as well as the few characters in the remaining ranges, is used for the IPA.

### Accented Letters

Accented letters used in Czech, Polish, and Slovak orthography are fairly straightforward: They are found in the Latin-1 Supplement and Latin Extended-A ranges of Unicode. But note that there are two different ways of representing accented letters in Unicode. Weslalex uses the *precomposed* method: < á> is a single character (U+00E1). There is another method, whereby one represents < á> as an <a> character followed by a <ˊ> character. Don’t use that representation. Searching for <aˊ> when Weslalex has < á> probably won’t work in Excel.

### International Phonetic Alphabet

In the past, people entered the IPA into the computer by actually inputting different characters, but using a font that represented those characters in unexpected ways. For example, to enter a <ʃ>, one scheme was to actually enter a <S>, but switch to a font that displays <S> as <ʃ>. Unfortunately, that scheme breaks down as soon as somebody changes or loses the font. If you have old files that were produced that way, you will find that when you copy and paste them into Weslalex files, they will transform back into the true base character: here, <S>. In general, if you have a file of IPA that is not Unicode, it probably will not work in Weslalex.

Unlike orthography, most accented letters in IPA have to be represented as a sequence of letters. Thus [r̝] (the sound of *ř*) is actually an <r> followed by a < ̝ >. Similarly a [r̩] (*r* used as a vowel) is an <r> followed by a < ̩ >. Exceptionally, the Polish nasal vowels [ẽ] and [õ] are each single characters. IPA diacritics must follow, not precede, their base. An exception to that rule is that the diacritic < ͡ > , which ties together the elements of affricates like [t͡s], goes between the two letters.

### Displaying Extended Characters

In Excel 2007 on Vista, the computer will display these characters tolerably well without your needing to do anything. If you do see a problem, it will probably take one of two forms.

* Badly presented characters. The ordinary letters will be in an attractive font, but the special characters will be in a different font that clashes. This is a minor problem, but you may want to fix it if you will be spending any time at all working with the spreadsheets.
* Empty boxes where special characters should be. This makes the data unusable for most purposes.

Both problems stem from using as your main font in Excel a font that does not contain the character in question. Vista will search through your other fonts to find the needed character. If it succeeds, you will likely get a badly presented, but technically correct, letter form. If it fails, and no file on your system contains that character, you will get an empty box. In either event, the solution is to select a font that contains the required letters.

The spreadsheets are set up to use the following fonts:

* Times New Roman for the text and pronunciation fields, such as spell, lemma, align, pron, and syll.
* Courier New for the morpho and cv fields.
* Calibri for other fields.

Under Vista, the spreadsheet will likely come up with the right fonts. If somebody has changed them in your copy of the Weslalex file, those fonts should be on your system, and you should be able to reset them: Select the column, press the right mouse button, Select > Table Column Data; Home, Font.

If you are using an older operating system, you may not have the specified fonts, or fonts by those names may lack the required characters. Times New Roman did not pick up many of the required characters until version 5.01, which started shipping with Microsoft Windows Vista and Apple OS X Leopard, and costs about 30 USD for other systems. A free solution is to download fonts from <http://scripts.sil.org/>. The fonts Doulos SIL, Charis SIL, and Gentium should have all the required characters. To see how well the fonts work, try looking at an entry like <háček> /haːt͡ʃek/; háčeks and the placement of the tie-bar over the t͡ʃ are the biggest challenges for most fonts.

### Inputting Extended Characters

When you wish to search the file for words that have certain letters or phonemes, you will need to type the desired characters into a search box. If these are special characters like <š> or <ʃ>, these may not be on your keyboard. There are several ways to work around such a problem.

**Copy and paste.** The simplest solution is to find the character somewhere, select it, copy it with Ctrl c, then paste it with Ctrl v. Table 1 below is one possible source from which to copy characters; it has all the characters used in Weslalex. A more specialized approach would be to make yourself a little cheat sheet containing just the characters you will be using regularly and that are not on your keyboard. A more general approach would be to use Web pages like <http://rishida.net/scripts/pickers/latin/> or <http://rishida.net/scripts/pickers/ipa/>, which have many additional characters you might like to use in other projects.

**Insert Symbol.** Excel comes with a tool that is handy for occasionally inserting special symbols. Under Insert, click on Symbol (the icon is an Ω). Up comes a window that lets you enter any character. Select the Symbols tab, and a relatively comprehensive font like Times New Roman. Click on a character to see its name at the bottom of the tool window. To insert it into Excel, double click the character, or hit the Insert button after selecting it. Insert Symbol keeps a handy list of the previous 20 characters you inserted so you don’t have to search for them again.

Fonts may have thousands of characters, so you may find it easier to find your character by first selecting the appropriate range from the Subset pulldown menu. The ranges are the same as the headers in Table 1.

**Input languages.** Vista provides a way to redefine what characters are produced when you depress certain keys on your keyboard. These are particularly useful for typing Czech, Polish, and Slovak. If you do not already have the ability to type characters such as <ů> from your keyboard, do the following to add an input language:

* Open the Control Panel
* Change keyboards or other input methods
* Keyboards and Languages tab
* Change keyboards ...
* Add
* Expand menu for a specific language: Czech, Polish, or Slovak
* Expand Keyboard
* Select a keyboard layout
* OK

You can install as many keyboards as you like. Switch between them by selecting the language in the Language Bar, or by pressing Shift Left Alt.

Vista does not come with any such keyboards for the IPA. Some have been developed by third parties, such as [Richard Collins](http://www.rejc2.co.uk/ipakeyboard/). If you wish to give it a try, download <http://www.rejc2.co.uk/ipakeyboard/IPA-Keyboard.zip> onto your desktop; open it and extra the files; then double-click setup.exe. Your language bar in Vista should now have an entry for English (United Kingdom) with keyboard United Kingdom IPA. When you select that keyboard, several keys will now generate IPA characters, as shown in Table 2. This keyboard is able to generate all the characters needed for Weslalex except for the Polish vowels [ẽ] and [õ].

You can also write your own keyboard layout. [The Microsoft Keyboard Layout Creator](http://msdn.microsoft.com/en-us/goglobal/bb964665.aspx) is a free tool for designing keyboards.

# The Weslalex Corpus

The corpus consists of the following children’s school books in the West Slavic languages—Czech, Slovak, and Polish.

Czech (387,702 tokens; 63,939 distinct word forms, 23,990 lemmas)

### Grade 1 (32,740 tokens; 8,608 distinct word forms, 4,362 lemmas)

Slabikář / Jiří Žáček, [ilustrovala] Helena Zmatlíková. – Vyd. 8., upravené. – Všeň : Alter, 2004. ISBN: 80-7245-063-8 (9,618 tokens)

Pracovní sešit ke Slabikáři, 1. díl / Hana Staudková a kol.– Všeň : Alter, 2004. ISBN: 80-7245-038-7 (3,669 tokens)

Pracovní sešit ke Slabikáři, 2. díl / Hana Staudková a kol. – Vyd. 1.– Všeň : Alter, 2004. ISBN: 80-7245-039-5 (4,675 tokens)

Čítanka pro prvňáčky / Jarmila Wagnerová. – 1. vyd.– Praha : SPN, 2003. ISBN: 80-7235-221-0 (3,832 tokens)

Učíme se číst : učebnice čtení pro 1. ročník základní školy ; zpracováno podle osnov vzdělávacího programu Základní škola s využitím genetické metody / Jarmila Wagnerová ; ilustroval Miloš Noll. – 1. vyd.– Praha : SPN, 2003. ISBN: 80-7235-000-5 (9,431 tokens)

Učíme se číst : pracovní sešit k 1. dílu učebnice: pro žáky 1. ročníku ZŠ / Jarmila Wagnerová, Jaroslava Václavovičová. – 1. vyd.– Praha : SPN, 2002. ISBN: 80-7235-196-6 (1,515 tokens)

### Grade 2 (57,144 tokens; 15,250 distinct word forms, 7,309 lemmas)

Čítanka 2 / Z. Nováková ; [ilustrovala] D. Wagnerová. – Vyd. 2. – Všeň : Alter, 2005. Half-title: Čítanka pro druhý ročník. ISBN: 80-7245-016-6 (28,393 tokens)

Čítanka pro 2. ročník základní školy: knížkake čtení, zpívání, hraní a malování / Josef Brukner, Miroslava Čížková.– 2. upravené vyd. – Praha : SPN, 2003. ISBN: 80-7235-222-9 (28,751 tokens)

### Grade 3 (77,322 tokens; 19,909 distinct word forms, 9,491 lemmas)

Čítanka 3 / [Leuka Bradáčová a kol.]. – Vyd. 1.– Všeň : Alter, 1997. Half-title: Čítanka pro třetí ročník. ISBN: 80-85775-67-0 (36,025 tokens)

Pracovní sešit k Čítance 3 / Miroslav Špika, Hana Standková.– Vyd. 1. – Všeň : Alter, 2004. (8,247 tokens)

Čítanka pro 3. ročník základní školy : knížka ke čtení, zpívání, hraní a malování / Josef Brukner, Miroslava Čížková ; ilustroval Miloš Noll.– 1. vyd. – Praha: SPN, 2000. ISBN: 80-85937-45-X (33,050 tokens)

### Grade 4 (106,176 tokens; 25,930 distinct word forms, 11,870 lemmas)

Čítanka 4 / [zpracoval kolektiv pod vedením Hany Rezutkové]. – Všeň : Alter, 1997. Half-title: Čítanka pro čtvrtý ročník. ISBN: 80-85775-49-2, 80-85775-69-7 (36,957 tokens)

Pracovní sešit k Čítance 4, první díl / [Miroslava Horáčková, Hana Staudková]. – Vyd. 1.– Všeň : Alter, 2003. ISBN: 80-7245-042-5 (8,633 tokens)

Pracovní sešit k Čítance 4, druhý díl / [Miroslava Horáčková, Hana Staudková]. – Vyd. 1.– Všeň : Alter, 2003. ISBN: 80-7245-043-3 (8,426 tokens)

Čítanka pro 4. ročník základní školy : knížka ke čtení, zpívání hraní a malování / Josef Brukner, Miroslava Čížková, Drahomíra Králová. – Praha : SPN, 2004. ISBN: 80-7235-263-6 (52,160 tokens)

### Grade 5 (114,320 tokens; 30,722 distinct word forms, 13,754 lemmas)

Čítanka 5 / [zpracoval kolektiv pod vedením Hany Rezutkové]. – Vyd. 2. – Všeň : Alter, 2003. Half title: Čítanka pro pátý ročník. ISBN: 80-85775-52-2, 80-85775-95-6 (45,988 tokens)

Pracovní sešit k Čítance 5, první díl / [Miroslav Špika, Hana Staudková]. – Vyd. 1.– Všeň : Alter, 2000 (8,320 tokens)

Pracovní sešit k Čítance 5, druhý díl / [Miroslav Špika, Hana Staudková]. – Vyd. 1.– Všeň : Alter, 2000 (8,225 tokens)

Čítanka pro 5. ročník základní školy : knížka ke čtení, zpívání, hraní a malování / Josef Brukner, Eva Beránková, Miroslava Čížková, Drahomíra Králová. – Praha: SPN, 1997. ISBN: 80-85937-71-9 (51,787 tokens)

## Polish (175,094 tokens; 34,361 distinct word forms)

### Grade 0 (6,478 tokens; 3,044 distinct word forms)

Mój kuferek : ćwiczenia dla sześciolatka, część 1 / Aleksandra Boniecka, Aleksandra Kozyra, Mirosława Wypchło.– Wyd. 2. – Warszawa: JUKA, 2005. ISBN: 83-7253-457-8 (1,498 tokens)

Mój kuferek : ćwiczenia dla sześciolatka, część 2 / Aleksandra Boniecka, Aleksandra Kozyra, Mirosława Wypchło. – Wyd. 2.– Warszawa: JUKA, 2005. ISBN: 83-7253-458-6 (1,426 tokens)

Mój kuferek : ćwiczenia dla sześciolatka, część 3 / Aleksandra Boniecka, Aleksandra Kozyra, Mirosława Wypchło. – Wyd. 2. – Warszawa: JUKA, 2005. ISBN: 83-7253-459-4 (1,548 tokens)

Mój kuferek : ćwiczenia dla sześciolatka, część 4 / Aleksandra Boniecka, Aleksandra Kozyra, Mirosława Wypchło.– [Wyd. 2?].– [Warszawa]: JUKA, [2005]? ISBN: 83-7253-478-0 (1,043 tokens)

Mój kuferek : ćwiczenia dla sześciolatka, część 5 / Aleksandra Boniecka, Aleksandra Kozyra, Mirosława Wypchło.– Wyd. 2. – Warszawa: JUKA, 2005. ISBN: 83-7253-479-9 (963 tokens)

### Grade 1 (28,010 tokens; 9,085 distinct word forms)

Świat ucznia: podręcznik do kształcenia zintegrowanego, klasa 1, część 1 / Barbara Mazur, Małgorzata Wiązowska, Katarzyna Zagórska. – Wyd. 1.– Warszawa: JUKA, 2004. ISBN: 83-7253-482-9 (10,662 tokens)

Świat ucznia: podręcznik do kształcenia zintegrowanego, klasa 1, część 2 / Barbara Mazur, Katarzyna Zagórska. – Wyd. 1.– Warszawa: JUKA, 2004. ISBN: 83-7253-488-8 (17,348 tokens)

### Grade 2 (57,411 tokens; 15,628 distinct word forms)

Świat ucznia: podręcznik do kształcenia zintegrowanego, klasa 2, część 1/ Katarzyna Grodzka. – Wyd. 1. – Warszawa: JUKA, 2003. ISBN: 83-7253-400-4 (27,719 tokens)

Świat ucznia : podręcznik do kształcenia zintegrowanego, klasa 2, część 2 / Katarzyna Grodzka.– Wyd. 1. – Warszawa: JUKA, 2003. ISBN: 83-7253-401-2 (29,692 tokens)

### Grade 3 (83,195 tokens; 22,237 distinct word forms)

Świat ucznia : podręcznik do kształcenia zintegrowanego, klasa 3, część 1 / Katarzyna Grodzka. – Wyd. 1. – Warszawa: JUKA, 2004. ISBN: 83-7253-412-8 (38,573 tokens)

Świat ucznia : podręcznik do kształcenia zintegrowanego, klasa 3, część 2 / Katarzyna Grodzka. – Wyd. 1.– Warszawa: JUKA, 2004. ISBN: 83-7253-413-6 (44,622 tokens)

## Slovak (180,577 tokens; 35,105 distinct word forms, 14,746 lemmas)

### Grade 1 (16,231 tokens; 6,399 distinct word forms, 3,627 lemmas)

Čítanka pre 1. ročník základných škôl / Lýdia Virgovičová.– 1. uprav. vyd. – Bratislava : Orbis Pictus Istropolitana, 2004. ISBN: 80-7158-495-9 (11,290 tokens)

Šlabikár pre 1. ročník základných škôl, I. časť / Lýdia Virgovičová. – 8. vyd.– Bratislava : Slovenské pedagogické nakladateľstvo, 1998. ISBN: 80-08-02841-6 (1,186 tokens)

Šlabikár pre 1. ročník základných škôl, II. časť / Lýdia Virgovičová. – 5. vyd.– Bratislava : Slovenské pedagogické nakladateľstvo, 1994. ISBN: 80-08-00295-6 (3,755 tokens)

### Grade 2 (41,767 tokens; 11,876 distinct word forms, 5,969 lemmas)

Čítanka pre 2. ročník základných škôl / Soňa Benková, Helena Komlóssyová, Jozef Pavlovič, Kamila Štefeková. – 5. vyd. – Bratislava : OG Poľana, 2003. ISBN: 80-89002-71-4 (23,646 tokens)

Slovenský jazyk pre 2. ročník základných škôl / Veronika Bakalová, Jarmila Krajčovičová, Anton Bujalka. – 4. vyd. – Bratislava : OG Poľana, 2004. ISBN: 80-89002-87-0 (18,121 tokens)

### Grade 3 (54,029 tokens; 14,996 distinct word forms, 7,169 lemmas)

Čítanka pre 3. ročník základných škôl / Soňa Benková, Helena Komlóssyová, Jozef Pavlovič, Kamila Štefeková. – 4. vyd. – Bratislava : OG Poľana, 2005. ISBN: 80-89192-14-9 (30,719 tokens)

Slovenský jazyk pre 3. ročník základných škôl / Veronika Bakalová, Jarmila Krajčovičová, Anton Bujalka. – 5. vyd. – Bratislava : OG Poľana, 2005. ISBN: 80-89192-12-2 (23,310 tokens)

### Grade 4 (68,550 tokens; 18,956 distinct word forms, 9,028 lemmas)

Čítanka pre 4. ročník základných škôl / Soňa Benková, Helena Komlóssyová, Jozef Pavlovič. – 3. vyd. – Bratislava : OG Poľana, 2005. ISBN: 80-89192-15-7 (40,097 tokens)

Slovenský jazyk pre 4. ročník základných škôl / Veronika Bakalová, Jarmila Krajčovičová, Anton Bujalka. – 4. vyd.– Bratislava : OG Poľana, 2005. ISBN: 80-89192-13-099999 (28,453 tokens)

The corpus consists of virtually all words found in these books. Excluded are:

* Tokens identified as not actually being words in the book’s main language.
* Instructions addressed to the teacher, in books intended for 6-year-olds (Grade 1 for Czech and Slovak, Grade 0 in Polish).
* Tokens that include characters other than letters.

# Organization of the Word Lists

Each language is presented in its own file: Czech in ces\_wf.xlsx, Polish in pol\_wf.xlsx, and Slovak in slk\_wf.xls. (The first three letters of those file names are the international standard [ISO 639-3](http://www.sil.org/iso639-3/) three-letter abbreviations for the language names.)

Sort Order

Each worksheet is alphabetized according to the usual conventions for the language in question. This section briefly summarizes the most striking differences from default European sorting rules.

### Czech and Slovak Sorting

In Czech and Slovak, <ch> is sorted as a single letter, which comes after <h>. Thus *hýbat* comes before *chápat*.

The letters <č>, <ř>, <š>, and <ž> are considered separate letters, whose alphabetical position immediately follows the corresponding letter that doesn’t have a háček. Thus *cvrk* comes before *čaj*.

Diacritics on other letters are usually ignored for sorting purposes. However, if two words would otherwise have the same sort position, diacritics are used as tie-breakers. A letter with a diacritic sorts after a letter without a diacritic. Thus *Česka* sorts before *česká*.

### Polish Sorting

In Polish, each letter with a diacritic sorts after the corresponding letter lacking a diacritic. Thus *oznaczać* comes before *ósmy*. <ź> comes before <ż>.

### Re-sorting

You can always sort a worksheet according to your own preferences. If you click in the first column and select Sort from the right-mouse-button menu, you will find the option to resort the data into the order that your own version of Excel and Vista thinks is natural. The advantage to this is that you yourself may find that order more natural too. The disadvantage is that if you tinker with the sorting, you may not be able to get back the correct Czech–Slovak or Polish order (except of course by Undo or loading a fresh copy of the spreadsheet).

## Parts of the File

The Czech and Slovak files each comprise two worksheets. The bulk of each spreadsheet is an Excel 2007 table. *Table* in this context is an unfortunately underspecified technical Excel term for a special type of layout that has rather more power than you get from just typing data in columns in a spreadsheet. Underneath the table are several rows that provide basic summary statistics for the rows visible in the table.

 All of the spreadsheets contain a worksheet called wf, which contains data for each wordform. It has a separate row for each inflected form or variant spelling in the corpus. In addition, the Czech and Slovak spreadsheets also contain a second worksheet, called lemma, which contains frequency information for individual lemmas.

## Wordform Data Fields

Each row in the data or wf table gives information about one *wordform type* found in the corpus. Our operational definition of a wordform is that two tokens that have the same spelling (ignoring case), belong to the same lemma, and have the same morphosyntactic analysis are the same wordform.

The discussion below walks you through each of the data fields, or columns, of the spreadsheets.

### spell

The spelling is given in the first column, which is labelled spell. All words are converted to lowercase, so that the distinction between upper- and lowercase is neutralized. Thus the row that begins **abeceda** collapses together information about words that are spelled “abeceda,” “Abeceda,” or “ABECEDA.” The entry **abraham** appears in lowercase here even though in practice it is always capitalized.

All accented letters in this column are precomposed. E.g., <č> is the single Unicode character U+010D, not the sequence U+0063 U+030C.

### lemma

In the Czech and Slovak spreadsheets, the second column indicates the lemma. A lemma is a word in the broad sense of the term: a lexical form, abstracting away from its inflection. E.g., the wordforms **abeceda**, **abecedě**, **abecedou**, **abecedu**, and **abecedy** are all forms of the same lemma. For convenience, lemmas are cited by a specific inflected form: for nouns, the nominative singular; for adjectives, the masculine nominative singular; for verbs, the infinitive. This is the form that appears in the lemma column. But it should be kept in mind that the lemma is actually a broader, more abstract entity, which comprises all inflected forms.

Occasionally two identical spellings (spell cells) are not merged into one, but occupy two different rows, because they are actually forms of two different lemmas. For example, in the Czech file there is a spelling **stát** whose lemma is **stát-1\_^(státní\_útvar)** (i.e., ‘state’) and one whose lemma is **stát-2\_^(něco\_se\_přihodilo)** (i.e., ‘to happen’). That is, there are two different wordforms, both spelt *stát*, which are differentiated because they are members of different lemmas.

In the Czech spreadsheet, the lemma cell can contain quite a bit of information. It always includes the spelling of the citation form of the lemma. If wordforms of that lemma are normally capitalized (e.g., proper nouns like *Drijverová*), the lemma is capitalized; unlike the spell column, there can be uppercase. Additional information is taken from the Hajič tagger dictionary:

* If two lemmas exists where the citation form has the identical spellings in the dictionary, their spellings are followed by differentiating tags -1, -2, etc.
* If the word names a cardinal number, then the spelling is followed by ` plus that number as digits, e.g., deset`10
* Verbs may be followed by \_: plus a code telling their aspect:
	+ T imperfect: brodit\_:T
	+ W perfect: napřímit\_:W
* Nouns may be followed by \_; plus a code telling their semantic field:
	+ E ethnonym: Polák\_;E
	+ G toponym: Polsko\_;G
	+ H chemistry: uranium\_;H
	+ K corporate: NATO\_;K\_^(North\_Atlantic\_Treaty\_Organization)
	+ L natural science: vemeník\_;L
	+ R product: Fiat-2\_;R\_^(vozidlo)
	+ S surname (family name): Foglar\_;S
	+ U medicine: antibiotikum\_;U
	+ Y given name: Anton\_;Y
	+ b economy, finances: napoleondor\_;b
	+ c computers and electronics: link-1\_;c
	+ o color: červený-1\_;o
* Words may be followed by \_, and a usage advisory. Sometimes the lemma citation form has been “corrected” to the modern standard, and the note actually applies only to the spell cell.
	+ a archaic: cykl\_,a
	+ e expressive: kolínko\_,e
	+ h colloquial: áčko\_,h
	+ l slang, argot: ksicht\_,l
	+ n dialect: chachar\_,n
	+ s bookish: čaromoc\_,s
	+ v vulgar: bréca\_,v
	+ x outdated spelling or misspelling: balkón\_,x
* Words can be followed by \_^ then a miscellaneous note in parentheses. A special type of note is the derivational one, which begins with a \* plus a digit, optionally followed by additional letters. If one subtracts the indicated number of letters from the end of the word then adds the specified letters, one will get the word that the word in question was derived from. E.g., “polámaný\_^(\*2t)” is derived from the lemma “polámat”: take “polámaný”, subtract two letters, then add “t”.

The tagging system is not applied with rigorous consistency throughout the file, nor have the Weslalex editors made any effort to proof it. Therefore it should be taken with a grain of salt, and is perhaps best applied as an aid in finding a few examples of a certain type of word.

### morpho

The third column in the Czech and Slovak worksheets is labelled morpho. It contains additional coded information about the wordform, mostly of a morphosyntactic nature. The information in this column was generated by versions of the Hajič disambiguating tagger. It has not been edited by hand and therefore should be used with a certain amount of caution. In this column this information is presented in its canonical, 15-character form. The information is positional, that is, each column stands for a different kind of information, and the codes must be interpreted in connection with that particular information type. For example, an N in the first position means that the word class a noun; an N in the third position means that its gender is neuter.

The information in this column is repeated, using more mnemonic codes, in the columns labelled pos, subpos, gender, number, case, possgender, possnumber, person, tense, grade, negation, voice, and var. These correspond in order to the 15 positions in the morpho column, except that the 13th and 14th positions in the morpho column are always empty (-) and therefore are not given a column of their own. See below for documentation for the remaining 13 columns.

Two words may have the same spell and lemma, but if their morpho field is different, they will be considered two different wordforms and given two separate rows in the data or wf table. For example, Czech has two rows that both have spell **slovo** and lemma “slovo”, but their morpho fields are different because one is analysed as having nominative case (NNNS**1**-----A----), and one is analysed as having accusative case (NNNS**4**-----A----). The three fields discussed so far—spell, lemma, and morpho—are the three fields that, taken together, uniquely determine wordforms. All other fields simply give further information about a wordform.

### analysis

The Polish file has an analysis column instead of lemma and morpho columns. Płotnicki’s Waspell tagger was used, which produces quite a different format from the information used in the Czech and Slovak files. Instead of codes it uses short abbreviations for morphosyntactic categories. It is not a disambiguating tagger; all analyses of ambiguous wordforms are given, separated by a **|** character. Each analysis consists of the citation form of the lemma, followed by either a ? (unknown word) or its analysis within parentheses. The grammatical information is not split up into different columns.

The spell and analysis columns uniquely determine the wordform in the Polish file. However, because of lack of disambiguation, there are very few instances where two rows have the same spell data.

### Frequency Columns

The following columns give information as to how often a wordform appears in the corpus. These frequency statistics are counted 4 ways in each of the grades covered by the language corpus. The column names are a concatenation of g plus the grade number plus the counting method: F, D, U, or SFI. For example, g1U is the *U* statistic for the wordform, computed over the first grade corpus. In addition, the final set of frequency columns tells the overall statistics for the language corpus as a whole. These appear without as grade-level prefix. Thus the column in the slk-wf.xlsx spreadsheet that is labelled simply U gives the *U* statistics for the words computed over the entire Slovak corpus.

The grades differ between the three corpora. Polish begins with g0—reception year—while Czech and Slovak begin with g1; all these correspond to 6 years of age. The corpora variously go up to g3 (Polish), g4 (Slovak), or g5 (Czech).

The ***F*** statistic tells how many times the wordform appears.

The ***D*** statistic tells the dispersion of the word across the grade or corpus. It is defined as

where *i* ranges over each book ID in the grade, is the probability of finding the word in that book (i.e., frequency of the word divided by the frequency of all tokens in the book), and *n* is the number of books in the grade. If the word has the same probability in each book, the dispersion *D* will be 1.0; if a word appears exclusively in one book, *D* will be 0. Situations between these extremes will have intermediate values. In the tables, *D* is reported to two decimal points, but precision up to four decimal points can be seen in the formula bar.

The ***U*** statistic is the estimated frequency per million tokens. Its formula is

As before, *F* is frequency and *D* is dispersion, and *i* ranges over each book ID in the grade. *N* is the number of all tokens in the corpus; is the frequency of the word in book *i*, and is the total number of tokens in that book. If the dispersion *D* is a perfect 1.0, the frequency is simply scaled up to a million. But that is adjusted downward the smaller the dispersion is. In the tables, *U* is reported as an integer, but four decimal positions are visible in the formula bar.

***SFI*** is the standard frequency index, which is simply a logarithmic transform of *U*:

This number is intended to give people a general feeling for how common a word is. In the tables, *SFI* is reported as an integer, but four decimal positions are visible in the formula bar.

If you are uncertain which of these measures to use, I would recommend *U*, in part because its meaning is relatively intuitive yet still generalizable: An estimate of how often the word would occur in a million-word text. It should be kept in mind, however, that it is a scaled estimate, and so one should guard against intemperate expressions such as saying that a word with a *U* of 5 occurred 5 times in the corpus.

### nlett

This field shows the number of letters in the spelling (content of the spell column). For the purpose of this statistic, diacritics are ignored, and digraphs such as <ch> are counted as two letters.

### Pronunciation

The next several columns deal with the pronunciation. The following principles were adhered to in all three languages:

* For each wordform, a single pronunciation is chosen.
* The pronunciation norm chosen is a very formal one that would typically be taught in schools as the standard pronunciation for the written literary language.
* The unit of transcription is the phoneme. That is, the transcription distinguishes between all sounds that can distinguish words within a language (phonemes). It intentionally avoids distinguishing sounds that cannot distinguish words (allophones). The working definition of a phoneme is a very traditional one, and does not take into consideration the possibility of archiphonemes, underspecification, derivation, phonological features, etc.
* Accordingly, stress is ignored, because it is essentially completely predictable given the segmental form of a word.
* Transcription uses the International Phonetic Alphabet.
* Transcription is as typographically simple as possible, while still being reasonably faithful to the pronunciation and expressing all phonemic contrasts. For example, the low vowel is transcribed as /a/ rather than the more precise, but phonemically otiose, [ä]. The mid front vowel is transcribed as /e/, even though the pronunciation in most words is closer to [ɛ].
* Affricates are transcribed in their full IPA glory with a tie bar, e.g. /t͡ʃ/, because sometimes affricates contrast with plosive + fricative sequences that would otherwise be transcribed the same way.
* Non-nuclear vocoids are consistently transcribed as glides, even in diphthongs, to make it clear that they do not begin or end their own syllable. E.g., *auto* is /awto/.

Because of the unavailability of machine-readable dictionaries, the pronunciations are generated by computer programs, which do not understand all the complexities and exceptions in the pronunciation system. If a pronunciation seems suspicious it may well be wrong, and should be corrected.

The following phonemic contrasts are symbolized by these IPA transcriptions:

|  |  |  |  |
| --- | --- | --- | --- |
| Phoneme | Czech | Slovak | Polish |
| a | **a**by | **a**by | **a**by |
| aː | d**á** | d**á** |  |
| b | **b**ez | **b**ez | **b**ez |
| c | a**ť**, dí**t**ě, **t**icho, poj**ď** | **ť**ažko, de**t**i, **t**icho | **k**im, **ki**edy |
| ɕ |  |  | **si**ada, **s**ilny, **ś**wiąt, we**ź** |
| d | **d**o | **d**o | **d**obra |
| d͡z | po**dz**imní | me**dz**i | **dz**wonek |
| d͡ʑ |  |  | **dzi**ało, **dź**więk |
| d͡ʒ | **dž**ungle | **dž**ungľa | **dż**ungla |
| e | b**e**z, člov**ě**k | b**e**z, p**ä**ť, člov**e**k | b**e**z |
| eː | dobr**é** | dobr**é** |  |
| ẽ |  |  | ci**ę**żki |
| f | **f**rancouzský , dí**v**ka, slo**v** | **f**arebne, **v**čera | **f**rancuski, bar**w** |
| ɡ | **g**azda, ni**k**de | **g**azda, ni**k**de | **g**rać, ni**g**dy |
| h | **h**las | **h**las |  |
| i | an**i**, b**y** | an**i**, b**y** | an**i** |
| iː | č**í**m, b**ý**vá | č**í**m, b**ý**va |  |
| ɨ |  |  | b**y** |
| j | **j**azyk, člov**ě**k, mů**j**, op**ě**t | **j**azyk, b**i**ely, mô**j** | **j**abłko, arm**i**a, b**i**erz |
| ɟ | **ď**ábel, **d**ělat, **d**ítě | **ď**alej, **d**ieťa, **d**eň | **g**imnazjum, **gi**ewoncie |
| k | **k**ůň, dialo**g** | **k**ôň | **k**ot, dialo**g** |
| l | **l**es | **l**es, vlk | **l**as |
| l̩ | v**l**k |  |  |
| l̩ː |  | st**ĺ**pcov |  |
| ʎ |  | učite**ľ** |  |
| m | **m**a**m**inka | **m**a**m**a | **m**a**m**a, d**ą**b |
| n | **n**a, číta**n**ka | **n**a, číta**n**ky | **n**a, choi**n**ka, b**ę**d**ą**c |
| ŋ |  |  | ci**ą**g |
| ɲ | aspo**ň**, **n**ě, **n**i, m**ě** | aspo**ň**, kô**ň**, **n**ebo, **n**ič | a**n**i, a**ni**oł, dło**ń** |
| o | **o**d, hlav**o**u | **o**d, hlav**o**u | **o**d |
| oː | g**ó**l | bal**ó**n |  |
| õ |  |  | bij**ą** |
| p | **p**án, chlé**b** | **p**án | **p**an, chle**b** |
| r | **r**ád | **r**ád | **r**ad |
| r̩ | s**r**dce | s**r**dce |  |
| r̩ː |  | k**ŕ**k |  |
| r̝ | **ř**íká, malí**ř** |  |  |
| s | **s**ám, be**z** | **s**ám | **s**am, bez |
| ʃ | **š**kola, a**ž** | **š**kola | **sz**koła, bie**rz** |
| t | **t**ak, doku**d** | **t**ak | **t**ak, doką**d** |
| t͡s | **c**elý | **c**elý | **c**ebula |
| t͡ʃ | **č**as, poněva**dž** | **č**as | **cz**as |
| t͡ɕ |  |  | łó**dź** |
| u | **u**čitel | **u**čiteľ | **u**cha, pag**ó**rki |
| uː | dol**ů, ú**lohy | **ú**lohy |  |
| v | **v**ás, d**v**ě | **v**ás | **w**as |
| w | cesto**u** | cesto**u**, pra**v**da, m**ô**že | **ł**ąka |
| x | **ch**ce, kni**h** | **ch**ce, kní**h** | **ch**ce, **h**alo |
| z | **z**a, Jo**s**ef | **z**a | **z**abawa |
| ʒ | **ž**ena | **ž**ena | **rz**adko |
| ʑ |  |  | **zi**arno, **z**ima, **ź**ródło |

In Slovak, devoicing is not always applied. Short /l̩/ is transcribed as /l/.

### pron

Presents the pronunciation as a simple string of phonemes, e.g., ʒviːkat͡ʃku

### syll

Presents the pronunciation, broken down into syllables. Each syllable is enclosed in angled brackets, e.g., <ʒviː><kat͡ʃ><ku> Currently the syllabification is based on simple phonetic principles. A single consonant between vowels goes with the second vowel, i.e., it forms the onset of a syllable with the following vowel. When consonants appear between vowels in a word, the last consonant goes with the next syllable. The consonant before last goes with the next syllable only if that makes a sequence of obstruent plus glide. The sounds /j/, /l/, /r/, / r̝/, and /v/ are treated as glides. Thus <a><kva><rel> but <dok><tor>. Morphology is not taken into account at all.

### nsyll

The number of syllables in the syll cell.

### nphon

The number of phonemes in the pron cell. Phonemes are counted as in the table above. Thus affricates like /t͡ʃ/ and long vowels like /uː/ are each treated as one phoneme. Diphthongs like /aw/ are treated as two phonemes each.

### cv

This is the syll cell, presented more abstractly: each consonant phoneme as a C, each vowel phoneme as a V. Thus <a><kva><rel> becomes <V><CCV><CVC>. The phoneme /w/ in Czech is treated as a vowel: *auto* /awto/ is <VC><CV>

### align

This field presents an alignment between spelling and pronunciation. This takes the form of a list of letter=sound correspondences, the correspondences separated from each other by a space. Correspondence is at the level of whole phonemes and whole letters. Almost always there is one letter to the left of the = sign and one phoneme to its right, but occasionally multiple letters spell one sound as a unit, e.g. Czech t=c i=i ch=x o=o ; and occasionally one letter spells multiple phonemes as a unit, e.g., Czech e=e x=ks k=k u=u r=r z=s

### Morphosyntactic Fields

In Czech and Slovak, the last 13 columns in the data table are the contents of the morpho field, expanded to make them more mnemonic. Note that all of these fields are generated automatically by versions of the Hajič tagger and have not been proofed by hand. In addition to the documented values, most fields may also contain the code - which means that the category is inapplicable to the lexeme in question. For example, nouns all have a - in the tense column.

### pos

Part of speech, or major word class. This column corresponds to the first character in the morpho field, but is longer and more memorable.

|  |  |  |
| --- | --- | --- |
| pos | morpho1 | Definition |
| adj | A | Adjective |
| adv | D | Adverb |
| conj | J | Conjunction |
| interj | I | Interjection |
| noun | N | Noun |
| other | X | Unknown |
| particle | T | Particle |
| prep | R | Preposition |
| num | C | Numeral |
| pron | P | Pronoun |
| verb | V | Verb |

### subpos

Detailed part of speech. This provides a more fine-grained view of the syntactic use of a word. The codes used in the morpho field (character 2) have the special restriction that any particular subpos code is always found with the same pos code, but that restriction is not carried over to this column.

|  |  |  |
| --- | --- | --- |
| subpos | morpho**2** | Definition |
| Codes used with pos = adj (adjectives) |
| hyph | 2 | Hyphenated |
| past-trans | M | Derived from verbal past transgressive form |
| poss | U | Possessive |
| pres-trans | G | Derived from present transgressive form of a verb |
| short | C | Nominal (short, participial) form |
| typical | A | General |
| Codes used with pos = adv (adverbs) |
| abs | b | Absolute (no negation or degrees of comparison) |
| grad | g | Graded (forming negation and comparison) |
| Codes used with pos = conj (conjunctions) |
| coord | ^ | Coordinating |
| subord | , | Subordinating (incl. *aby*, *kdyby* in all forms) |
| Codes used with pos = interj (interjections) |
| typical | I | Interjections |
| Codes used with pos = noun |
| typical | N | General |
| Codes used with pos = num (numbers) |
| card>4 | n | Cardinal ≥ 5 |
| card<5 | l | Cardinal, 1 through 4 |
| fract | y | Fraction ending in *-ina*, used as a noun |
| gen-1 | h | *(ne)jedny* |
| gen-adj | d | Generic with adjectival declension |
| gen-noun | j | Generic ≥ 4 used as a noun |
| gen-short | k | Generic ≥ 4 used as an adjective, short form  |
| indef | a | Indefinite |
| indef-adj | w | Indefinite, adjectival declension |
| mult | v | Multiplicative, definite |
| mult-indef | o | Multiplicative indefinite |
| mult-rog | u | *kolikrát* |
| ordin | r | Ordinal (adjective declension) |
| ordin-rog | z | *kolikátý* |
| rog | ? | *kolik* |
| times | \* | *krát* ‘times’ |
| Codes used with pos = particle |
| typical | T | Particle |
| Codes used with pos = prep (prepositions) |
| phras | F | Partial; only appears in a phrase |
| typical | R | General, without vocalization |
| vowel | V | Preposition, with vocalization -*e* or -*u* |
| Codes used with pos = pron (pronouns) |
| demon | D | Demonstrative |
| indef | Z | Indefinite |
| L | L | *všechen*, *sám* |
| n-pers | 5 | *on* ‘he’ after a preposition (with prefix *n*-) |
| n- rel- jenž | 9 | Relative *jenž*, *již*, ... after a preposition |
| neg | W | Negative |
| O | O | *svůj*, *nesvůj*, *tentam* |
| pers | P | Personal |
| pers-clit | H | Personal, enclitic (short) form |
| pers-poss | S | Possessive *můj*, *tvůj*, *jeho* |
| prep | 0 | Preposition + *ň*: *naň*, *proň*, etc. |
| refl-long | 6 | Reflexive *se* in long forms |
| refl-poss | 8 | Possessive reflexive *svůj* |
| refl-short | 7 | Reflexive *se*, *si* ± -*s* |
| rel-claus | E | Relative *což* |
| rel- jenž | J | Relative *jenž*, *již*, ... not after a preposition |
| rel-poss | 1 | Relative possessive |
| rel-rog | Q | Relative/interrogative *co*, *copak*, *cožpak* |
| rel-rog-adj | 4 | Relative/interrogative with adjectival declension |
| rel-rog-anim | K | Relative/interrogative *kdo* |
| rel-rog-encl | Y | Relative/interrogative *co* as an enclitic |
| Codes used with pos = verb |
| pres | B | Present or future form |
| pres-ť | t | Present or future tense, with the enclitic *-ť* |
| cond | c | Conditional (of the verb *být* only) |
| imper | i | Imperative |
| infin | f | Infinitive |
| part-past-act | p | Past participle, active |
| part-past-act-ť | q | Past participle, active, with the enclitic *-ť* |
| part-past-pass | s | Past participle, passive |
| trans-past | m | Transgressive past |
| trans-pres | e | Transgressive present (endings -*e*/-*ě*, -*íc*, -*íce*) |
| Codes used with pos = other |
| other | X | Not in dictionary |

### gender

|  |  |  |
| --- | --- | --- |
| gender | morpho3 | Definition |
| -a | Q | Feminine (with singular) or neuter (with plural) |
| any | X | Any |
| fem | F | Feminine |
| fem/neut | H | Feminine or neuter |
| mas | Y | Masculine |
| mas-anim | M | Masculine animate |
| mas-inan | I | Masculine inanimate |
| mas/neut | Z | Not feminine |
| neut | N | Neuter |
| -y | T | Masculine inanimate or feminine |

### number

|  |  |  |
| --- | --- | --- |
| number | morpho4 | Definition |
| -a | W | Singular for feminine gender, plural with neuter |
| any | X | Any |
| dual | D | Dual , e.g. *nohama* |
| plur | P | Plural, e.g. *nohami* |
| sing | S | Singular, e.g. *noha* |

### case

|  |  |  |
| --- | --- | --- |
| case | morpho5 | Definition |
| acc | 4 | Accusative |
| any | X | Any |
| dat | 3 | Dative |
| gen | 2 | Genitive |
| inst | 7 | Instrumental |
| loc | 6 | Locative |
| nom | 1 | Nominative |
| voc | 5 | Vocative |

### possgender

Gender of possessor:

|  |  |  |
| --- | --- | --- |
| possgender | morpho6 | Definition |
| any | X | Any |
| fem | F | Feminine |
| mas-anim | M | Masculine animate |
| mas-inanim |  |  |
| mas/neut | Z | Not feminine |

### possnumber

Number of possessor:

|  |  |  |
| --- | --- | --- |
| possnumber | morpho7 | Definition |
| plur | P | Plural |
| sing | S | Singular |
| any | X | Any |

### person

|  |  |  |
| --- | --- | --- |
| person | morpho8 | Definition |
| 1st | 1 | 1st person |
| 2nd | 2 | 2nd person |
| 3rd | 3 | 3rd person |
| any | X | Any person |

### tense

|  |  |  |
| --- | --- | --- |
| tense | morpho9 | Definition |
| any | X | Any |
| fut | F | Future |
| past | R | Past |
| pres | P | Present |

### grade

Comparison degree:

|  |  |  |
| --- | --- | --- |
| grade | morpho10 | Definition |
| comp | 2 | Comparative |
| pos | 1 | Positive |
| superl | 3 | Superlative |

### negation

|  |  |  |
| --- | --- | --- |
| negation | morpho11 | Definition |
| aff | A | Affirmative (not negated) |
| neg | N | Negated |

### voice

|  |  |  |
| --- | --- | --- |
| voice | morpho12 | Definition |
| act | A | Active |
| pass | P | Passive |

### var

Classification of word variant:

|  |  |  |
| --- | --- | --- |
| var | morpho15 | Definition |
| arch/coll | 3 | Very archaic, also archaic + colloquial |
| arch/lit | 4 | Very archaic or bookish, but standard at the time |
| coll | 6 | Colloquial (standard in spoken language) |
| coll/infreq | 7 | Colloquial (standard in spoken language), less frequent variant |
| infreq | 1 | Variant, second most used (less frequent), still standard |
| rare | 2 | Variant, rarely used, bookish, or archaic |
| special | 9 | Special uses |

## Lemma Worksheets

The lemma worksheets available in the Czech and Slovak spreadsheets are closely analogous to the wf worksheets, except that the main entries are lemmas, not wordforms. For example, the Slovak wf worksheet has entries for **abeceda**, **abecedy**, **abecedu**, and **abecede**, but the lemma worksheet only as one row comprising them all. That row is labelled **abeceda**, but the information it gives pertains to all the inflected forms, and **abeceda** is just a convenient label, as used in dictionaries. Therefore the lemma worksheet does not attempt to give pronunciations, word lengths, and so forth. The main reason it exists is to present count statistics that apply to the lemma as a whole. The *F* (raw count) fields for a given lemma like **abeceda** could be straightforwardly derived by summing the *F* statistics for each of the lemma’s wordforms in the wf worksheet. But the *D* (dispersion) statistics are derived from counting how often a wordform – or lemma – appears in each of the books, and so lemma statistics cannot be straightforwardly derived from the wordform statistics. Indeed, the dispersion for a lemma will be higher than that of its individual wordforms, if it appears in multiple wordforms. The *D* of the individual wordforms **abeceda**, **abecedy**, etc., range from 0 to .70, but the *D* of the lemma as a whole is .72. Because *U* and *SFI* are based on *D*, the lemma-wise versions of those counts must likewise be looked up in the lemma worksheet.

## Column Summaries

The rows that follow the table provide summary statistics for the values in the columns above them. Number of Rows tells how many rows (wordforms) appear in the table. The other data summarize the numbers in the column below which they appear. The rows Minimum and Maximum tell the range; Mean is the ordinary average, and appears in a darker colour to highlight its importance; and Standard Deviation is the standard deviation.

The empty row with a dark blue background, immediately after the table data (but above the column summaries section), provides a hook for doing other types of statistics. Left-click in one of those cells to get a handle for a pull-down menu offering other statistical options. A particularly useful option is Sum, but several sophisticated functions are also available.

## Filtering Data

It is fairly easy to select wordforms that have specific values in one or more of the columns.

Perhaps the most familiar approach is to use Ctrl f (viz., Home > Find & Select > Find). This tool will find and highlight the value you are looking for. For example, to find words that have subpos value “indef” (indefinite), one would select the subpos column, do Ctrl f, type “indef” in Find what, and click Find All. You will get a scrollable list of addresses that have that value; clicking on an address will take you to that datum and highlight it. Be aware that this process can take a surprisingly long time.

Another approach is to use table filters. These are fast and convenient for many kinds of word searches. Ctrl Shift L (viz., Data > Filter, the funnel icon) turns on or off the ability to filter data. When filtering is enabled, all the column heads (the ones with the names, not the ones with the Excel column designators like AA) have a pulldown menu handle. The menu provides several ways of searching the data. In all cases, when you search the data, rows that do not match your search criteria become invisible. You see only the rows the match the criteria. Of course they are still there, behind the scenes, and all rows keep their original row numbers.

For example, to find interjections, you can search for wordforms that have interj in the pos field. Go to the top of the table, pull down the menu for pos, and you will see two different ways to search for interj. In this case, since there are only a dozen or so different values in this column, the easiest approach is to look at the list of values and make sure that only the box with interj is checked. Note that all values are probably checked. You could manually uncheck every value except the desired interj. Or, you could uncheck the (Select All) box, which means deselect all, then check the interj: 2 clicks instead of 10. Click OK, then notice that the table is quite a bit shorter, because you will see only the words that the tagger decided were interjections.

Note that the label for the pos field now has a funnel icon to show that the table is filtered by this column; if you hover the mouse over this, a tooltip will tell you that it is filtering for “interj”. Note also that the row numbers are unchanged. Finally, you will notice that all the summary statistics after the table have been updated, to reflect only the visible rows. However, you should be aware that if you write your own Excel formulas to search or otherwise manipulate the table, they will typically look at all the data, not just the visible data. Be sure you understand the rules.

You may filter by multiple fields simultaneously. For example, to find only the CVC interjections, keep the previous filter in force, and use the drop down menu on the cv field to select <cvc>.

On text fields, such as those containing spellings and pronunciations, the filter options let you do fairly sophisticated things such as searching for cells that begin with certain values (good, e.g., for finding prefixes) or end with certain values (e.g., suffixes), or contain certain values. For example, to find wordforms where <t> spells /c/, one can filter by the align field, “Text filter contains t=c” The text searches are case-insensitive, e.g., “a” is treated the same as “A”. In Weslalex, this is seldom a problem, but be on guard against certain situations where case is important, most notably the lemma and morpho fields.

On numeric fields, such as the frequency counts, you will find several useful options, such as the ability to search for numbers above or below a certain value, or above or below the mean.

There is a special menu option for clearing a filter, i.e., making it stop hiding columns. A quicker way to turn off filtering is to type Ctrl Shift L (or click the big Data funnel icon).

Of course, Excel provides many other ways to sort, search, and otherwise process data. Rather than add hundreds of pages to this short guide, I refer the reader to Excel’s fine documentation.

Table 1. *Characters Arranged by Unicode Order*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Basic Latin |  | Latin-1 Supplement |  | Latin Extended-A |  | IPA Extensions |
| ! | U+0021 |  | Á | U+00C1 |  | ą | U+0105 |  | ɕ | U+0255 |
| " | U+0022 |  | É | U+00C9 |  | ć | U+0107 |  | ɟ | U+025F |
| ' | U+0027 |  | Í | U+00CD |  | Č | U+010C |  | ɡ | U+0261 |
| ( | U+0028 |  | Ó | U+00D3 |  | č | U+010D |  | ɨ | U+0268 |
| ) | U+0029 |  | Ô | U+00D4 |  | Ď | U+010E |  | ɲ | U+0272 |
| \* | U+002A |  | Ú | U+00DA |  | ď | U+010F |  | ʃ | U+0283 |
| + | U+002B |  | Ý | U+00DD |  | ę | U+0119 |  | ʎ | U+028E |
| , | U+002C |  | á | U+00E1 |  | ě | U+011B |  | ʑ | U+0291 |
| - | U+002D |  | ä | U+00E4 |  | Ĺ | U+0139 |  | ʒ | U+0292 |
| . | U+002E |  | é | U+00E9 |  | ĺ | U+013A |  | Spacing Modifier Letters |
| / | U+002F |  | í | U+00ED |  | Ľ | U+013D |  | ː | U+02D0 |
| 0  | U+0030  |  | ó | U+00F3 |  | ľ | U+013E |  | Combining Diacritical Marks |
| – | – |  | ô | U+00F4 |  | Ł | U+0141 |  | ̝ | U+031D |
| 9 | U+0039 |  | õ | U+00F5 |  | ł | U+0142 |  | ̩ | U+0329 |
| : | U+003A |  | ö | U+00F6 |  | ń | U+0144 |  | ͡ | U+0361 |
| ; | U+003B |  | ú | U+00FA |  | Ň | U+0147 |  | Latin Extended Additional |
| < | U+003C |  | ü | U+00FC |  | ň | U+0148 |  | ẽ | U+1EBD |
| = | U+003D |  | ý | U+00FD |  | ŋ | U+014B |  |  |
| > | U+003E |  |  |  |  | ŕ | U+0155 |  |  |  |
| ? | U+003F |  |  |  |  | Ř | U+0158 |  |  |  |
| A  | U+0041  |  |  |  |  | ř | U+0159 |  |  |  |
| – | – |  |  |  |  | Ś | U+015A |  |  |  |
| Z | U+005A |  |  |  |  | ś | U+015B |  |  |  |
| [ | U+005B |  |  |  |  | Š | U+0160 |  |  |  |
| ] | U+005D |  |  |  |  | š | U+0161 |  |  |  |
| ^ | U+005E |  |  |  |  | Ť | U+0164 |  |  |  |
| \_ | U+005F |  |  |  |  | ť | U+0165 |  |  |  |
| ` | U+0060 |  |  |  |  | Ů | U+016E |  |  |  |
| a | U+0061 |  |  |  |  | ů | U+016F |  |  |  |
| – | – |  |  |  |  | ź | U+017A |  |  |  |
| z | U+007A |  |  |  |  | Ż | U+017B |  |  |  |
| | | U+007C |  |  |  |  | ż | U+017C |  |  |  |
|  |  |  |  |  | Ž | U+017D |  |  |
|  |  |  |  |  |  | ž | U+017E |  |  |  |

Table 2. *Entering IPA Characters With the Collins Virtual Keyboard*

|  |  |
| --- | --- |
| For IPA, | Type |
| ŋ | M |
| ɕ | &c |
| ɟ | -j |
| ɡ | g |
| ɨ | -i |
| ɲ | <n |
| ʃ | S |
| ʎ | @y or "y  |
| ʑ | &z |
| ʒ | Z |
| aː | a: |
| r̝ | r\_r |
| r̩ | r\_' |
| t͡s  | t^=s |

*Note.* Input is actually defined by the location of the key on your keyboard, not by the key’s label. The inputs were designed for a British keyboard; labels on other keyboards may differ.