Analýza staročeské morfologie v Excelu

BORIS LEHEČKA, BORIS@DALIBORIS.CZ
ODDĚLENÍ VÝVOJE JAZYKA
ÚSTAV PRO JAZYK ČESKÝ AV ČR

LINGVISTIKA PRAHA 2014
11. DUBNA
16.00

Boris Lehečka

- boris@daliboris.cz
- oddělení vývoje jazyka ÚJČ AV ČR
- programátor s lingvistickými základy
- Vokabulář webový <http://vokabular.ujc.cas.cz>
- materiály ke stažení
  - http://vokabular.ujc.cas.cz/informace.aspx?t=LP2014
  - http://bit.ly/1qmOjzD

Obsah

- Účel analýzy
  - deklinace staročeských apelativ
  - časové období
  - vzory
- Auditorium
  - anketa
- Excel
  - Power Query
    - import a transformace datových zdrojů
    - programovací jazyk

Obsah

- Excel
  - PowerPivot
    - stamiliony položek
    - tabulky
    - relace
  - Kontingenční tabulky a grafy
- Vstupy
  - výchozí
  - po transformaci
    - pomůcky
- Ukázka

Účel analýzy

- deklinace staročeských apelativ
  - disertační práce Pavlíny Jínové
- časové období
- vzory
  - lemmata
  - koncovky
- ověření výskytu tvaru/tvarů

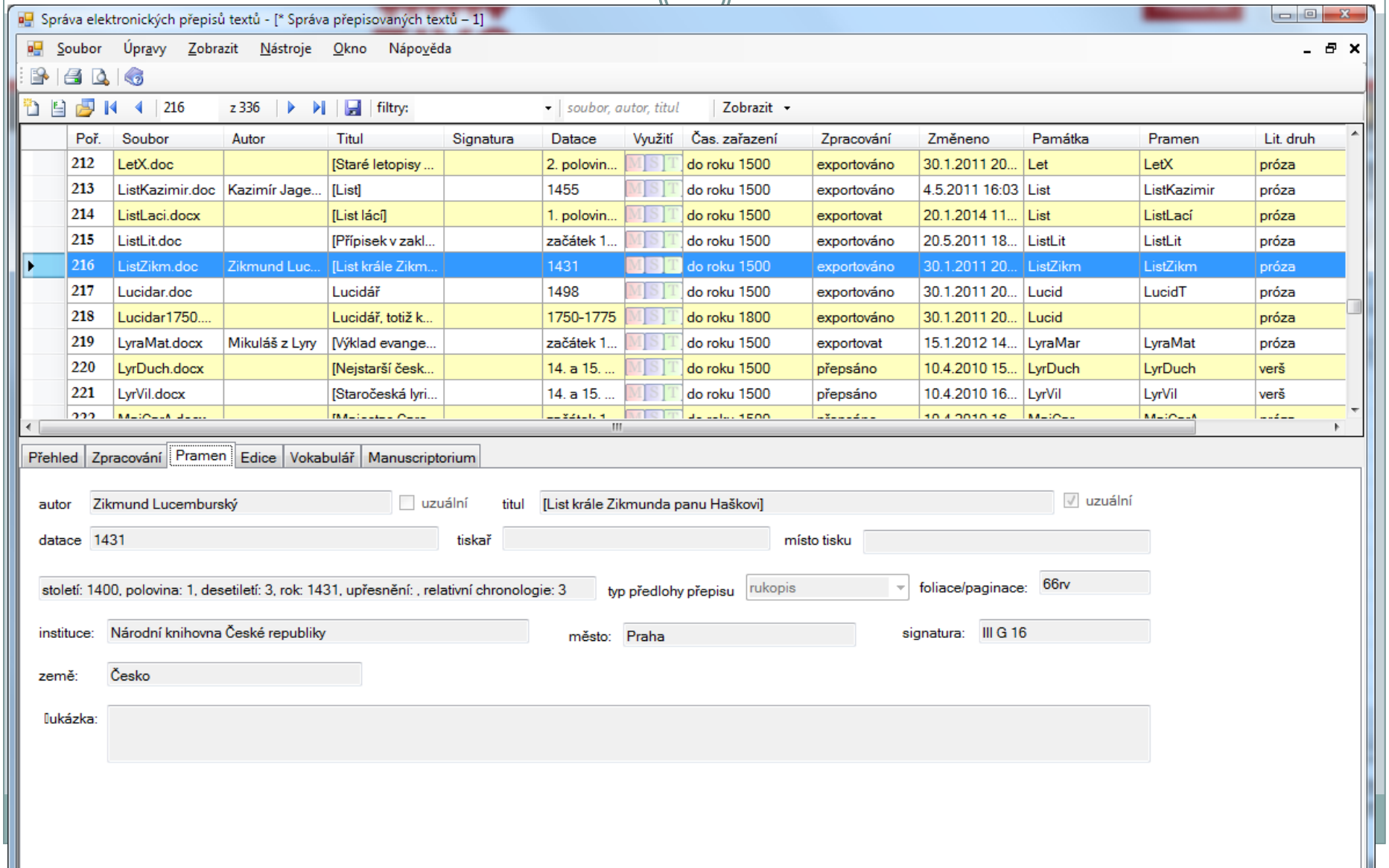
Anketa

- Kdo používá Excel?
- Jakou verzi Excelu?
  - 2010
  - 2013
  - jinou (např. Office 365 pro vysokoškoláky)
- Jakou edici Excelu?
  - Home and Student
  - Professional
- Kdo zná PowerPivot?
- Kdo zná Power Query?
- Kdo zná vertikálu?
- Příprava dat, nebo kontingenční tabulky/grafy v Excelu?

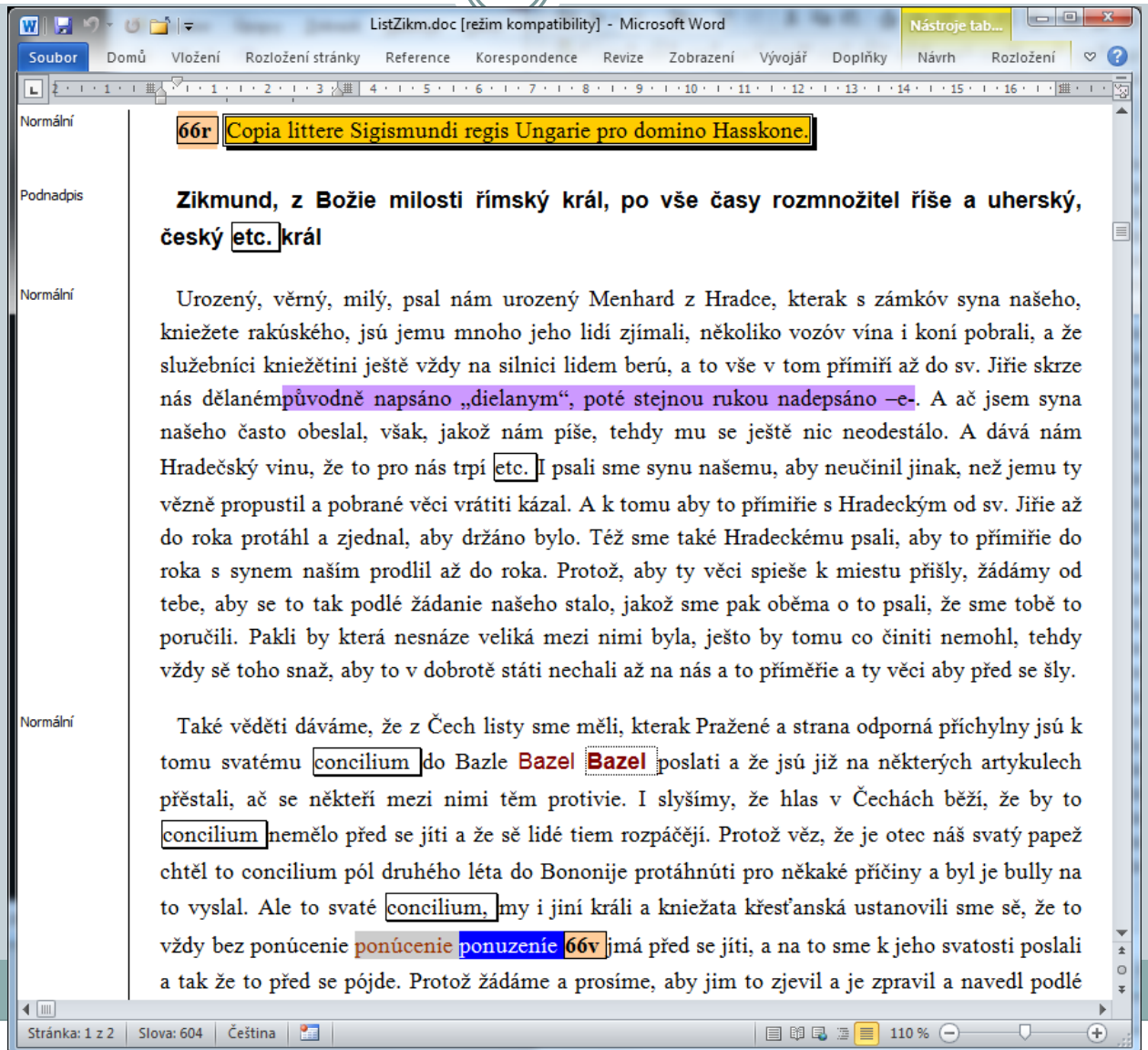
Vstupy

- Staročeská textová banka
  - Metadata
    - identifikátor
    - zkratka
    - období vzniku
    - literární žánr
    - atp. (podle potřeby)
  - Vertikála
    - poznámky = metainformace

Staročeská textová banka – metadata



Staročeská textová banka – Word



Vstupy

- Slovníky
  - ESSČ (*Přib–ž, ž–ch*)
  - MSS (*a–ž*)
  - StčS (*n–při*)
  - GbSlov (*a–netbanlivý*)
  - formát XML

Úpravy vstupů

- Vertikála
  - kategorizace tokenů
    - jazyk
    - torzo
    - funkce (interpunkce, číslo)
    - relevance
  - segmentace tokenů na fonogramy
    - zakončení
      - 1–3 fonogramy

Fonogram

- grafická jednotka korespondující s fonémem
  - změna fonému v důsledku hláskoslovného vývoje, flexe a/nebo slovotvorby znamená změnu fonogramu

m	**ú**	**ch**	**a**
m	**ú**	**š**	**ě**
m	**ou**	**š**	**e**

d	**ie**	**v**	**k**	**a**
d	**í**	**v**	**k**	**a**

h	**o**	**s**	**t**	**ie**
h	**o**	**s**	**t**	**í**

h	**o**	**s**	**t**	**i**	**e**
h	**o**	**s**	**t**	**i**	**í**

Úpravy vstupů

- Metadata o textech
  - výběr údajů
  - datace
    - časová období
  - převod z XML na CSV
- Slovníky
  - lemma
  - slovní druh
  - morfologická charakteristika

Úpravy vstupů – nástroje

- C#
  - transformace
    - DOCX > XML
    - XML > VERT
      - VERT > TAB
    - Metadata > TAB
- OpenRefine
  - http://openrefine.org
  - analýza vstupů
    - statistiky
    - redefinice vlastností

Excel – Power Query

- Doplněk
  - verze 2.10.3598.81
  - pro Excel 2010 a 2013
  - podmínky
    - Windows Vista až Window 8.1
    - Office 2010 Professional Plus + SA
    - Office 2013 Professional Plus, Office 365 ProPlus, Excel 2013
- Princip
  - import dat z různých zdrojů
  - nový programovací jazyk
  - není třeba řešit zabezpečení maker

Excel – PowerPivot

- Microsoft SQL Server 2012 PowerPivot for Excel
- Doplněk
  - pro Excel 2010
  - od Excelu 2013 je již součástí programu
  - verze 11.0.3129.0
- Podmínky
  - Windows XP SP3 až Window 8
  - Office 2010 zdarma pro všechny edice
  - v Excelu 2013 návrh pouze v edici Professional
- 64bitová verze (pro velké objemy dat)

Excel – PowerPivot

- Princip
  - stamiliony záznamů
  - tabulky
  - relace
- Nevýhody
  - nerozlišuje velikost písmen
  - relace
    - M : N
    - text != jedinečný klíč
  - textová data
    - velký objem
    - bez redukce

Excel – PowerPivot

- Kontingenční tabulky
- Kontingenční grafy
- Průřezy

Excel – doporučení

- Vstupní data
  - textové soubory
    - rychlé generování
    - lze i v Excelu
    - identifikátory
      - relace
  - tabulky v sešitě
    - ad hoc metadata
  - aktualizace
    - v PowerPivotu
    - v Excelu
- Import v PowerPivotu
  - nejprve nastavit parametry, pak vybrat soubor
    - první řádek jako názvy, až po výběru souboru

Ukázka

- Metadata
  - XML
  - TXT
  - úpravy
- Textová banka
  - Word
  - XML TEI P5
  - Vertikála

Ukázka

- PowerPivot
  - import dat
  - vytvoření relací
  - počítané sloupce
  - míry
  - testovací data
    - malý objem
- kontingenční tabulka
  - vytváření
  - interaktivita
  - průřezy
    - pro více objektů
- kontingenční graf
  - vytváření
  - interaktivita
  - průřezy

Ukázka

- Power Query
  - vytvoření dotazu
  - opětovné spuštění dotazu
- Sešit Tokeny
  - reálná data
  - připravené tabulky
  - připravené grafy
  - tabulka a graf na přání

Analýza morfologie v Excelu

- Plusy
  - off-line
  - není třeba korpusový manažer
  - opakovatelnost (s jinými daty)
  - ad hoc analýzy
  - kontingenční přehledy
  - zdarma (pro Excel 2010)
- Minusy
  - příprava vstupních dat
  - chybí kontext
  - bez pokročilých analýz

Postupy

- Míry
  - Formátování čísel
  - Velikost písmen
- Identifikátory
- Aktualizace
  - PowerPivot
  - Kontingenční tabulky
  - Propojené tabulky
  - Power Query
  - Soubory
- Průřezy