Úvod do praxe stínového řečníka

Příprava jazykových modelů

Jazykový model - trénování

- získání textů
- čištění (nechat jen to, co se má rozpoznávat)
- tokenizace (oddělení rozpoznávacích jednotek)
- normalizace (převod čísel, zkratek, nestandardních slov atd.)
- unifikace (sjednocení synonym, multislova atd.)
- výpočet pravděpodobností n-gramů
- míchání jazykových modelů z různých zdrojů

Získání textů

- koupí textových korpusů
- z internetu
  - zpravodajské servery (ČTK, Parlamentní listy, iDNES, …)
  - titulky (iVysílání, OpenSubtitles, …)
  - diskuzní fóra
- od zákazníků
  - medicínské databáze
  - advokátní dokumenty
  - soudní rozhodnutí
- přepisem zvukových záznamů

Čištění textu

- ponechat jen text, který by se rozpoznával
- specifické pro různé zdroje a formát dat

=== 29. 142093
=== 30. 142094
čas vyš.10:40Subjektivně:~Objektivně:~Dg:~Doporučení:~Kontrola:~fdghrtyhrtw
=== 31. 142099
Odběry hormonů štítné žlázy - vše v normě, bez patologického nálezu.
=== 32. 142100
Subjektivně:~ padání vlasů nadále, má obavy z počítáníObjektivně:stav stale stejný bez výrazného zhoršení , tr, test poz. Pacientka nyní aplikuje asi 2 měsíce Regain, zlepšení nepozoruje. Neocapil netolerovala. Bere navíc potravinové doplňky ale rovněž bez efektu. Trpí častými IMC, nyní bude řešeno na urologii, kde nevyloučen ani zakrák chirurgický. Dg:Alopecia areataDoporučení:Rp. Panthenol inj., lo amp, Ibal., lxtý i.m., 2. Methionin cps lOO,O, 2xl, Framykoin ung, I bal., l-2xdenně - eroze na předloktí. Dále zítra ráno odběry šž.Kontrola:za 2-3 měsíceKódy výk onů:44013,44239 , 44261
=== 33. 142101
sKre: 85, sUre: 5,0, sCB: 66,0, sAST: 0,63, sALT: 0,55, sBil: 10,0, sALP: 1,50, sCho: 4,20, sKM: 202, sGMT: 0,17, sNa: 144, sK: 4,8, sCl: 108

Tokenizace textu

- oddělení (popř. spojení) rozpoznávacích jednotek
  - : ; % " ? ! / ( ) + = [ ] * § $ { }
  - pomlčky, spojovníky (2-3, vědecko-technický, …)
  - čárky (657 354,30)
  - tečky (657.354, 55., J.P., atd., …)

Dg:Alopecia areata
Doporučení:Rp. Panthenol inj., lo amp, Ibal., lxtý i.m., 2. Methionin cps lOO,O, 2xl, Framykoin ung, I bal., l-2xdenně - eroze na předloktí.

"Přehlídka byla slavnostně zahájena v Ostravě a poputuje po 20-30 městech České republiky. Celkem promítneme 101 soutěžních a dva nesoutěžní filmy z šestnácti zemí světa," uvedla 2. pořadatelka Eva Kadlecová.

Normalizace textu

- rozvinutí číselných údajů
  - v češtině mluvnické kategorie - pád, osoba, číslo
  - Part-Of-Speech (POS) tagging – morfologické značky

" Přehlídka byla slavnostně zahájena v Ostravě a poputuje po 20 - 30 městech České republiky . Celkem promítneme 101 soutěžních a dva nesoutěžní filmy z šestnácti zemí světa , " uvedla 2. pořadatelka Eva Kadlecová .

Dne 13. 10. 1997 bylo usnesením č . j . ORHK – 1895 / TČ - 80 - 2006 zahájeno trestní stíhání. Na základě smlouvy o půjčce půjčil můj klient dne 11. 11. 1996 paní Marii Novákové částku 125600 korun českých . Marie Nováková se zavázala tuto částku splatit do 6. 10. 1975 s 8 % úrokem z prodlení ročně . Tuto smlouvu jste podepsala jako 1. ručitel dlužníka a zavázala se dlužnou částku splatit na účet vedený u Komerční banky a . s . , ve Vysokém Mýtě .

Unifikace textu

- dekapitalizace velkých písmen na začátcích vět
- náhrady podle slovníků (zavedení multislov)

Dne třináctého desátý tisíc devět set devadesát sedm bylo usnesením **č . j .** ORHK – tisíc osm set devadesát pět / TČ - osmdesát - dva tisíce šest zahájeno trestní stíhání. *Na* základě smlouvy o půjčce půjčil můj klient dne jedenáctého jedenáctý tisíc devět set devadesát šest paní Marii Novákové částku sto dvacet pět tisíc šest set **korun českých** . *Marie* Nováková se zavázala tuto částku splatit do šestého desátý tisíc devět set sedmdesát pět s osmi % úrokem z prodlení ročně . *Tuto* smlouvu jste podepsala jako první ručitel dlužníka a zavázala se dlužnou částku splatit na účet vedený u **Komerční banky a . s .** , ve **Vysokém Mýtě** .

Po špatném **French Open** a špatném Wimbledonu odehrál **Roger Federer** řekněme solidní sérii turnajů na americké půdě , semifinále **indianapolis** , čtvrtfinále kanadského **mistroství** , druhé kolo **sincinety**, kde zvítězil až v **tie breaku** .

Náhradové slovníky

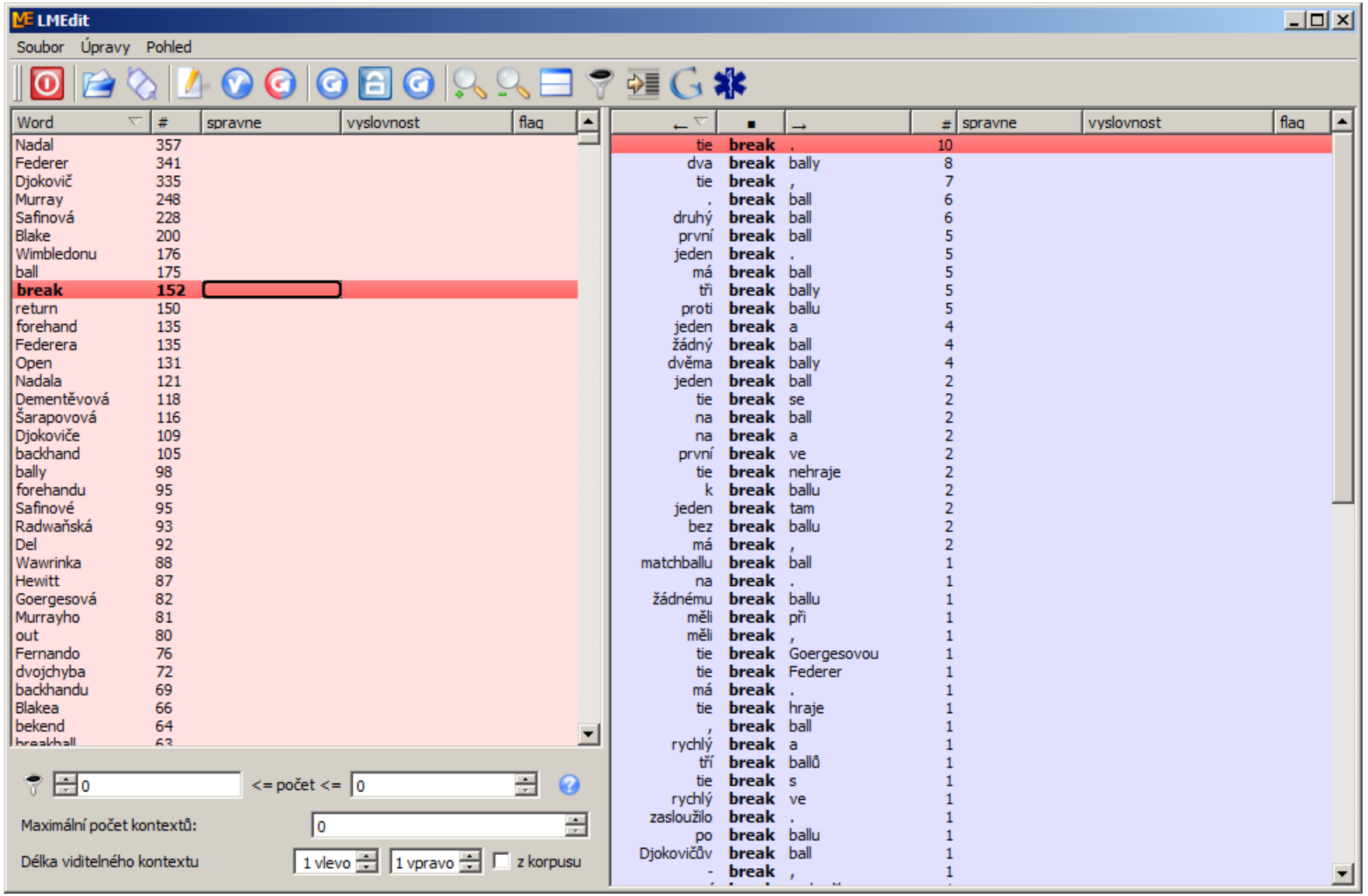
- specifické slovníky pro medicínu, advokacii, sport atd.
- třísloupcový slovník
  - 1. sloupec – originální text (původní zápis)
  - 2. sloupec – správný text (výsledný zápis)
  - 3. sloupec – výslovnost(i)

a propos	a_propos	apropó
ačkoli	ačkoliv	ačkoli;ačkoliv
ccm	cm3	cé cé em;cé em tři; centymetr krychlový;centymetry krychlové;…; kubický centymetr;kubické centymetry;…
ČT24	ČT_24	čé té dvacet čtyři
Západočeské univerzitě	Západočeské_univerzitě	západočeské unyverzytě
Zanzibar	Zanzibar	
Zanzibarem	Zanzibarem	

Stávající slovníky

- běžná česká slova – 3 miliony položek
- příjmení v ČR – 952 tisíc položek
- názvy firem registrovaných v ČR – 341 tisíc firem
- názvy obcí a ulic v ČR – 145 tisíc položek
- ostatní slova, názvy apod. – 16 tisíc položek
- křestní jména v ČR – 10 tisíc položek
- názvy států, národností, jazyků a velkých měst – 10 tisíc položek
- sportovní výrazy – 5 tisíc položek
- čísla – 590 položek
- interpunkční znaménka a příkazy – 20 položek

Program LMEdit



Zpracování slovníku v programu LMEdit

Návod na zpracování slovníku v programu LMEdit

Slovník v programu LMEdit představuje jednotlivá slova ze zdrojových textů, která nejsou obsažena v již existujících slovnících. Tato slova je tedy potřeba zpracovat, tzn. označit jako správná, opravit chybná slova, popř. jim nadefinovat výslovnosti. Takto zpracovaná slova se přidají do již existujících slovníků, pomocí kterých se opět zpracují zdrojové texty.

Program LMEdit

V levé části programu LMEdit jsou zobrazeny jednotlivé položky slovníku a jejich četnost, tedy kolikrát se daná položka vyskytla ve zdrojových textech. Položky se budou zpracovávat od těch nejvíce četných postupně k méně četným. Seřadit podle četnosti lze položky kliknutím na název sloupce "#" vzestupně či sestupně.

V pravé části programu jsou zobrazeny kontexty (tedy okolí daného slova), ve kterých se slovo vyskytlo. U jednotlivých kontextů je opět jejich četnost a opět je dobré tyto kontexty seřadit podle četnosti kliknutím na název sloupce "#" vzestupně či sestupně. Velikost levého a pravého kontextu (počet okolních slov) lze měnit v okénku "Délka viditelného kontextu" v levé části dole (je potřeba odtrhnout "z korpusu").

Provádění oprav

Při zpracování slovníku je potřeba dbát na velikost písmen.

Při zpracování každé položky slovníku může nastat jedna z těchto možností:

1) Slovo existuje výhradně samostatně a má svůj vlastní význam (musí být zřejmé, o co jde)

Slovo ve slovníkové části zkopírujte do sloupce "spravne" (je možno dvojklikem do sloupce a zmáčknutím SHIFT+F1). Pokud slovo v LMEditu není správně, ale je zřejmý jeho správný zápis, ve sloupci "spravne" uveďte tento správný zápis. Do sloupce "vyslovnost" zapište výslovnost, pokud je potřeba (viz dále).

Hovorové výrazy je potřeba zapsat v gramaticky správném tvaru, přičemž do výslovnosti se uvede jak výslovnost odpovídající nespisovné variantě (pokud nejde o překlep nebo přeřek), tak výslovnost spisovného tvaru (pokud nejsou stejné). V případě, že nelze hovorový výraz nahradit stejně dlouhým výrazem spisovným, nechá se hovorový, popř. jeho spisovnější verze.

Příklady:

mistrině	mistryně	
zalehá	zaléhá	zaléhá;zalehá
starovním	startovním	
biathlonistu	biatlonistu	bijatlonystu
jedenačtyřicet	jednačtyřicet	jednačtyřicet;jedenačtyřicet
vyhovujou	vyhovují	vyhovují;vyhovujou
Polský	polský	
přemotivovaný	přemotivovaný	přemotyvovaný
servismani	servismani	servismani;servismeni
světákách	světácích	světákách;světácích

2) Slovo existuje pouze ve spojení s jiným výrazem (jinými výrazy) a tvoří tak běžně používaný významový celek

V kontextové části slova najděte časté výrazy, se kterými zpracovávané slovo tvoří významové celky. K tomu použijte omezení viditelného kontextu slova tak, aby na řádce zůstal pouze tento významový celek. Do sloupce "flag" poté zapište "m". Pokud výraz v LMEditu není správně, ale je zřejmý jeho správný zápis, do sloupce "spravne" uveďte tento správný zápis. Totéž platí, pokud je potřeba spojit např. chybně rozdělené slovo. Do sloupce "vyslovnost" zapište výslovnost, pokud je potřeba (viz dále). Kromě ustálených slovních spojení (v češtině např. "de facto") jde zejména o názvy (geografické, sportovních klubů atd.) a (cizí) jména, tedy o významové celky začínající velkým písmenem.

Příklady:

Lake Placid	lejk plesid	m

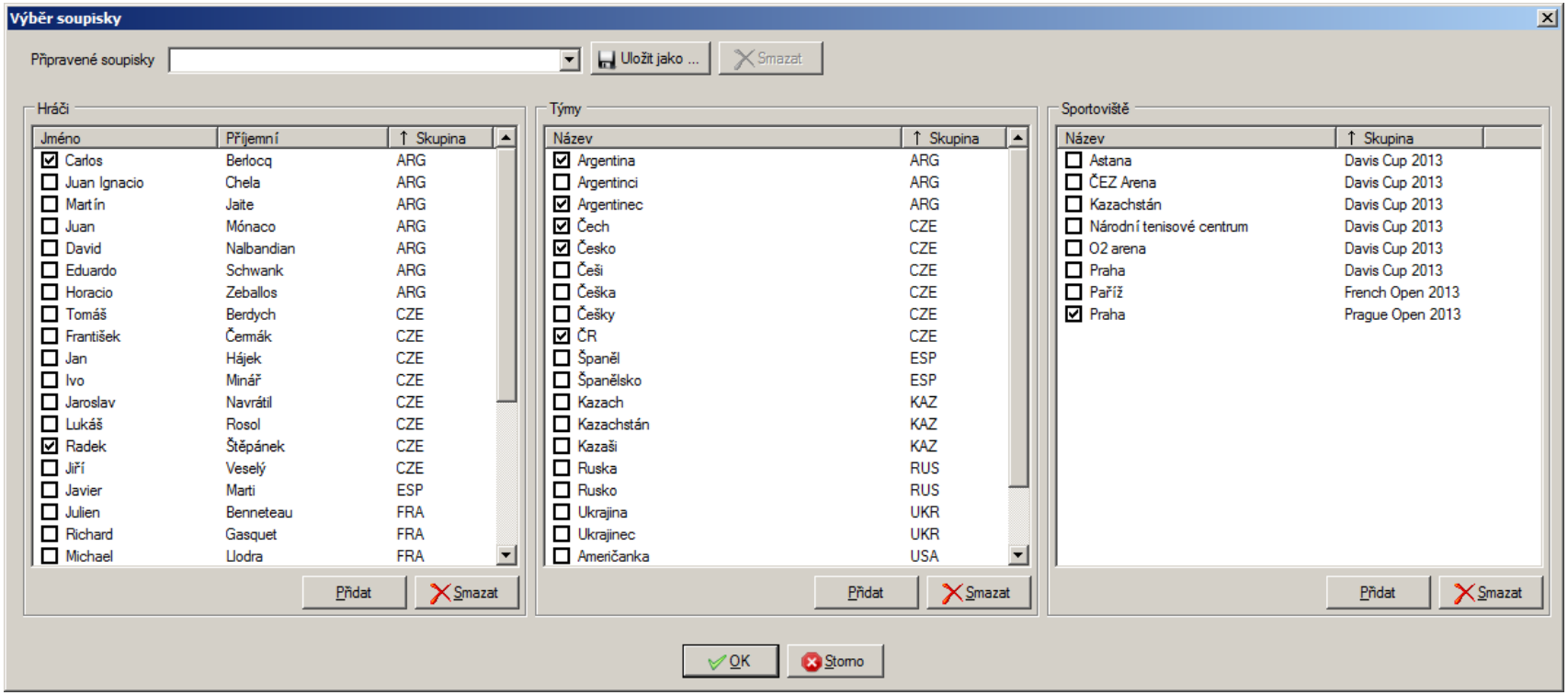
Sportovní soupisky

- sportovní texty obsahují jména a příjmení hráčů, názvy sportovních klubů nebo národností, popř. názvy stadionů, kurtů atd., které jsou často specifické pro konkrétní zápas (fotbal, hokej apod.)
- označení jazykových tříd
  - jména sportovců
  - názvy států, klubů, národností
  - názvy sportovišť (státy, města, stadiony)

spousta zajímavých událostí se děje na ledě [haly Jubilejnyj 2] , kde sledujeme utkání {Česko 1} {Rusko 1}, ve kterém září (Jaromír Jágr 1) s {ruským 7} útočníkem (Ovečkinem 7)

Sportovní soupisky

- třídový n-gramový jazykový model
  - 12 tříd pro jména sportovců
  - 12 tříd pro názvy států, klubů, národností
  - 6 tříd pro názvy sportovišť (států, měst, stadionů)



Míchání jazykových modelů

- na základě vzorového textu jsou automaticky určeny váhy jednotlivých modelů pro minimalizaci perplexity (složitosti) úlohy
  - HOKEJ
    - zápas – 55 % hokej, 23 % tisk, 17 % MF, 5 % TVR
    - studio – 21 % hokej, 32 % tisk, 32 % MF, 15 % TVR
  - FOTBAL
    - zápas – 50 % fotbal, 25 % tisk, 20 % MF, 5 % TVR
    - studio – 28 % fotbal, 21 % tisk, 34 % MF, 17 % TVR
  - TENIS
    - 63 % tenis, 22 % tisk, 10 % MF, 5 % TVR

Příprava jazykových modelů - shrnutí

- problém sběru velkého množství kvalitních dat z cílové domény (a jejich čištění)
- proces zpracování dat pro jazykové modely je mírně závislý na zdroji dat, doméně, jazyku apod.
- pokud má být cílový slovník jazykově čistý, je potřeba ruční zpracování dat (náhradové slovníky apod.)
- speciální domény (sport, parlament apod.) vyžadují expertní přístup
- je za tím spousta (nekonečné) práce