VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA PARALELNÍHO KORPUSU Z TITULKŮ K FILMŮM

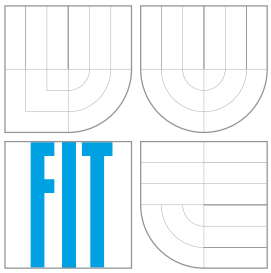
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

MAREK STRAŇÁK

BRNO 2009

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA PARALELNÍHO KORPUSU Z TITULKŮ K FILMŮM

AUTOMATIC CREATION OF PARALLEL CORPUS FROM MOVIE SUBTITLES

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — MAREK STRAŇÁK
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2009

Abstrakt

Táto práca sa zaoberá tvorbou paralelného korpusu, ktorého zdrojom sú filmové titulky. Konkrétne sa jedná o zarovnanie českých a anglických viet s využitím slovníkov a morfologických analyzátorov, prípadne zarovnanie titulkov v iných jazykoch na základe časovania jednotlivých komentárov. Práca taktiež pojednáva o obecnej problematike paralelných korpusov.

Abstract

This work is about the creation of parallel corpus, where movie subtitles is main source. In particulary, it is about alignment czech and english sentences using dictionaries and morphologic analyzers or alignment talks of subtitles in other languages using timing of talks. The work give basic information about parallel corpus.

Klíčová slova

paralelný korpus, titulky, zarovnanie, značkovanie

Keywords

parallel corpus, subtitles, alignment, tagging

Citace

Marek Straňák: Automatická tvorba paralelního korpusu z titulků k filmům, bakalářská práce, Brno, FIT VUT v Brně, 2009

Automatická tvorba paralelního korpusu z titulků k filmům

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana doc. RNDr. Pavla Smrža, Ph.D.

.....
Marek Straňák
17. mája 2009

Poděkování

Moje poďakovanie patrí doc. RNDr. Pavlovi Smržovi, Ph.D. za jeho odbornú pomoc a trpezlivosť pri vypracovaní bakalárskej práce. Ďalej by som chcel poďakovať kamarátovi Františku Svobodovi, od ktorého som prevzal databázu s titulkami.

Obsah

Kapitola 1

Úvod

Písmo má svoju vlastnú históriu. Na jeho počiatku boli rôzne jaskynné maľby, ktoré vyjadrovali snahu človeka zaznamenať svoje myšlienky a pocity. Tieto piktogramy boli jednoduché obrázky, ktoré znázorňovali celé vety, a nie samostatné znaky a slová ako to je v súčasnosti. Vznik dnešnej podoby písma bol podmienený u národov, ktoré dospeli na taký stupeň vývoja, že jeho potreba bola nevyhnutná a prvé písomné práce môžeme nazvať akýmisi základmi korpusov. Každá kultúra však dosiahla rôzneho stupňa písma. Jeho vývoj ovplyvňovala zemepisná poloha, spoločenské potreby, osobitnosť jazyka a taktiež písmo susedných národov. Každý jazyk má svoje zákonitosti, čo sa prelínalo do písma. Začali vznikať prvé slovníky a k nim potrebné korpusy. Niektoré jazyky majú presné pravidlá pre stavbu vety ako je tu u angličtiny, iné však spôsobujú značné problémy pri formalizácii daného jazyka, napríklad čeština. V mojej práci som sa snažil zrovnať práve tieto dva jazyky.

1.1 Cieľ práce

Cieľom mojej práce je pochopiť analýzu, značkovanie a spracovanie paralelných korpusov a jeden takýto korpus vytvoriť a rozšíriť tak súčasnú množinu korpusov. Konkrétne sa bude jednať o paralelný korpus zložený z filmových titulkov, kde si budú navzájom odpovedať české a anglické vety. Snažil som sa navrhnúť a vytvoriť jednoduchú, rýchlu a efektívnu metódu zarovnania titulkov. Na záver zhodnotím môj výsledný korpus a zhrniem jeho možné využitie.

Kapitola 2

Korpusy

Aby som sa mohol venovať rozboru korpusov, najprv sa pokúsim vysvetliť, čo to vlastne korpus je. Anglická encyklopédia [6] popisuje korpus nasledovne: *„In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe“*. Encyklopédia [1] definuje pojem korpus ako: *„rozsiahly súbor elektronicky uložených jazykových dát určených k vedeckému výskumu“*. O korpuse môžem povedať, že je to rozsiahly súbor textov, ktoré sú označkované a uložené najčastejšie v elektronickej podobe. Väčšinou sa používajú pre štúdium slov, ich významov a najčastejších kontextov. Používajú sa v rôznych oblastiach lingvistiky (od morfológie, cez syntaxiu, sémantiku až po štylistiku). Okrem klasických rozsiahlych korpusov existujú aj špecifické korpusy malého rozsahu (rádovo niekoľko stoviek tisíc slov), taktiež možno nájsť aj zvukové a vizuálne korpusy, a aj korpusy s neoznačkovaným textom.

2.1 Druhy korpusov

Korpusy možno rozdeliť podľa rôznych kritérií na niekoľko skupín, pričom jednotlivé skupiny sa môžu ďalej členiť do kategórií, ako je to uvedené v dokumente [24]. Jednotlivé korpusy sa však od seba odlišujú svojou veľkosťou, typom, jazykom, zdrojom textov, značkovaním, atď.

2.1.1 Delenie podľa média

Korpusy sa delia do dvoch základných skupín, a to na textové (písané) a zvukové (hovorené). Základnou zložkou u písaných korpusov sú rôzne texty, knihy, časopisy, scenára, ale aj prepisy hovorenej reči, či už bežnej konverzácie alebo rádiového a televízneho vysielania a rozhovorov. U hovorených korpusov sú to zvukové záznamy reči, najčastejšie magnetofónové nahrávky. V súčasnosti prevládajú hlavne textové korpusy. Hlavnou príčinou ich nadradenosti je obrovská investícia a úsilie, ktoré je potrebné vynaložiť na vytvorenie zvukového korpusu, ktoré sú mnohonásobne vyššie ako u písaných korpusov. Vznikajúce zvukové korpusy sú zvyčajne malé.

2.1.2 Delenie podľa časového obdobia

Ďalšie dôležité členenie korpusov je na synchronné a diachronné. Synchronný korpus je založený na textoch súčasného jazyka relatívne krátkeho časového obdobia (niekoľko posledných desaťročí), v priebehu ktorého môžeme považovať jazyk za nemenný systém. Tieto texty (korpusy) reprezentujú súčasný jazyk. Na druhej strane, diachronný korpus popisuje jazyk v rôznych vývojových fázach, poprípade celý jeho vývoj, a preto obsahuje texty z rozsiahlejšieho obdobia. Umožňuje sledovanie vývoja jednotlivých jazykových javov. Od synchronného korpusu sa odlišuje rôznymi vnútornými ale aj vonkajšími charakteristikami. V dnešnom svete prevládajú hlavne synchronné korpusy, a to vďaka veľkému množstvu dostupných zdrojov. Diachronné korpusy takmer vôbec neexistujú v zvukovej podobe a texty v starších vývojových stupňoch jazyka sú obmedzené a nevyrovnané vzhľadom na zastúpenie jednotlivých typov (prevažujú veršované a náboženské texty, legendy).

2.1.3 Delenie podľa obsahu

Vo všeobecnosti sa každý korpus zameriava na určitú skupinu textov. Podľa toho ich možno rozdeliť do nasledujúcich skupín:

- národné korpusy
- paralelný korpus (viacjazyčné korpusy)
- korpus nárečových textov
- korpus básnických textov
- korpus hovorených prejavov
- korpus lexikografických diel
- vzdelávací korpus (texty upravené pre štúdium)
- a iné

2.2 Príklady korpusov

Brown Corpus of Standard American English (Brown Corpus) [9]

Patrí medzi najvýznamnejšie korpusy. Vznikol v roku 1964 na Brownovej univerzite a je pokladaný za prvý moderný elektronický korpus. Jeho zakladateľmi sú Henry Kučera a W. Nelson Francis. Obsahuje vždy približne 2 000 slov z 500 rôznych zdrojov. Všetky zdroje pochádzajú z roku 1961 a možno ich rozdeliť do 15tich kategórií. Celkovo korpus obsahuje 1 014 312 slov. V roku 1979 sa objavilo jeho druhé vydanie, ktoré obsahovalo značky (celkovo 81 druhov značiek). V súčasnosti je už zastaraný a patrí ku korpusom s menším rozsahom, ale aj napriek tomu sa stále používa.

BNC (British National Corpus) [7]

Vznikol v priebehu rokov 1991–1994 pod vedením Oxford University Press. Patrí medzi jednojazykové synchronné korpusy. Obsahuje široký prierez britskej angličtiny z konca 20. storočia. Obsahuje viac ako 100 miliónov slov. Približne 10% slov tvorí menší korpus hovoreného jazyka, ktorý pochádza z bežných neformálnych rozhovorov (nahrávaných dobrovoľníkmi rôzneho veku a z rôznych spoločenských vrstiev) až po verejné prejavy a rádiové

vysielania. Zvyšnú časť tvorí písaný jazyk, ktorý obsahuje texty z regionálnych novín, odborných periodík a časopisov, beletrie ale aj zo školských a univerzitných prác. Každý text zaradený do korpusu obsahuje úplne kontextové a bibliografické informácie. V roku 2001 vyšla druhá, upravená verzia pod názvom BNC Word. Išlo o opravu chýb nachádzajúcich sa v záhlaviach textov alebo v SGML značkovaní.

Český národný korpus (ČNK) [24]

Vlastný korpus ako celok neexistuje, ale skladá sa z niekoľko menších korpusov rôzneho rozsahu. Najväčší z nich je SYN2000, ktorý zachytáva súčasný český jazyk a obsahuje viac ako 100 miliónov slovných tvarov. Ide o synchronný korpus, ktorý je lemmatizovaný a morfologicky označený. Obsahuje texty z rôznych oblastí, 60% textov tvorí publicistika, 25% odborná literatúra a 15% beletria. Zo SYN2000 bol neskôr odvodený ďalší korpus PUBLIC, ktorý je verejne dostupný na internete. Jeho rozsah je však len okolo 30 miliónov slovných tvarov. Medzi ďalšie korpusy, ktoré tvoria ČNK, patria diachronný korpus DIAKORP (približne 1,75 miliónov tvarov), synchronné hovorené korpusy ORAL-PMK (Pražský mluvený korpus, 700 000 tvarov) a BMK (Brněnský mluvený korpus, 500 000 tvarov) a nárečový korpus DIAL (synchronný aj diachronný, 100 000 tvarov).

Kapitola 3

Paralelný korpus

Patrí do špeciálnej skupiny korpusov, tzv. viacjazyčné korpusy. Obsahujú nie len originálne texty, ale aj ich mutácie v rôznych jazykoch. Jednotlivé texty sa obsahovo i formálne podobajú a sú zarovnané tak, aby si jednotlivé časti navzájom odpovedali. To znamená, že pre každú jednoznačne určenú časť textu v originálnom jazyku je priradená časť textu v inom jazyku. Tieto časti môžu byť veľkosti odstavcov, alebo menších celkov ako sú vety alebo priamo slová a frázy.

Texty pre paralelné korpusy vznikajú najčastejšie prekladom z jedného jazyka do druhého. Paralelné korpusy však nemusia vždy obsahovať texty v dvoch čí viacerých jazykoch, ale môže sa jednať aj o rôzne verzie rovnakého textu v jednom jazyku (rôzne verzie prekladu).

Súhrne môžem povedať, že paralelný korpus obsahuje zrovnateľné dáta vo viacerých podobách, ktoré sa odlišujú jazykom alebo verziou prekladu. Tento druh korpusov sa stáva v poslednom čase čoraz viac aktuálny. Jeho hlavnou výhodou je rozsiahle obohatenie dvojjazyčných slovníkov. Od bežných korpusov sa odlišuje v spracovávaní dvoch alebo viacerých zrovnateľných textov, využívajú sa iné metódy a prostriedky. Každý z dvojice textov paralelného korpusu možno pokladať za samostatný korpus.

3.1 Využite

Zo všetkých korpusov v súčasnej dobe sa dostáva najväčšiemu využitiu práve paralelným korpusom. Ich použitie je všestranné. Pracujú s nimi nie len ľudia, ale sú používané aj na ďalšie spracovanie pomocou rôznych programov a aplikácií. Najväčšie uplatnenie nachádzajú u lexikografov a rôznych jazykových vedcov (porovnanie lexikálnej, morfologickej a syntaktickej štruktúry jazykov), prekladateľov ale aj učiteľov a študentov cudzích jazykov. Používajú sa na tvorbu dvojjazyčných slovníkov (párovanie slov, viacslovných výrazov), tvorbu synoným, interpretáciu textu v jednom jazyku na základe iného jazyka, vyhľadávanie informácií súčasne vo viacerých jazykoch, alebo automatický strojový preklad a kontrola prekladu.

3.2 Nevýhody

Hlavnou nevýhodou u paralelných korpusov je veľmi častý nesprávny preklad textov, mnohé nie sú autentické a majú rozdielnu štruktúru. Paralelne je možné získať len malú časť z celkového počtu textov, a aj to len niektoré typy. Predpokladané využitie dosiahnu len vtedy, ak budú zarovnané aspoň na úrovni viet, čo je niekedy komplikované. Používanie

automatických nástrojov vyžaduje zvyčajne špeciálne znalosti v danej oblasti a výsledky nie sú nikdy 100%-né (zarovnanie nie je vždy správne) a párovanie je potrebné ešte ručne opravovať.

3.3 Existujúce paralelné korpusy

Kačenka [3]

Názov vznikol ako akronym a znamená „Korpus anglicko-český, elektronický nástroj Katedry anglistiky“. V podstate je to paralelný korpus českého a anglického jazyka, kde jednotlivé texty sú zarovnané na úrovni viet. Bol vytvorený Filozofickou fakultou Masarykovej univerzity v spolupráci s Fakultou Informatiky v Brne v roku 1997 s úmyslom umožniť analýzu prekladu na ucelených textoch. Obsahuje bibliografické texty v anglickom jazyku a ich české preklady. Väčšina anglických textov pochádza z internetu, naopak české texty museli byť skenované. V súčasnosti obsahuje 3 297 283 slov, pričom práve 1 689 513 bolo získaných pomocou skenovania. Základnou myšlienkou kačenky je použitie, ktoré by nevyžadovalo špeciálne počítačové znalosti alebo korpusové manažéry. Preto sú jednotlivé texty uložené v jednoduchej forme. Celý korpus je uložený na jednom CD, ale jeho použitie je obmedzené autorským zákonom.

Intercorp [18]

Korpus Intercorp je projekt, ktorého cieľom je vybudovať rozsiahly paralelný synchronný korpus obsahujúci čo najviac jazykov. Na jeho tvorbe sa podieľali učitelia a študenti Filozofickej fakulty Univerzity Karlovej v Prahe a spolupracovníci ÚČNK (Ústav Českého národního korpusu). Celý projekt je akademický a nekomerčný. Obsahuje väčšinou zarovnané beletrické texty v češtine a v ďalších jazykoch, pričom čeština je akýsi základný jazyk. Každý text má jednu českú verziu (originál alebo preklad), ktorá je zarovnaná s textami v iných jazykoch. Niektoré sú opatrené morfologickými značkami. V súčasnosti je prístupných 505 textov v rôznych jazykoch (angličtina, francúzština, maďarčina, slovenčina, čeština, atď.). Celý korpus obsahuje približne 31 157 000 českých slov a 34 464 000 slov v iných jazykoch. Počet jazykov a spracovaných textov sa však neustále zvyšuje. Korpus je prístupný iba cez špeciálne rozhranie, ktoré je nadstavbou nad systémom Manatee.

Hansards [11]

Patrí medzi najznámejšie paralelné korpusy. Je to francúzsko-anglický paralelný korpus a skladá sa z diskusií kanadského parlamentu v 80.rokoch. Obsahuje viac ako 1,3 miliónov textov a 50 miliónov slov. Stal sa určitým štandardom pre paralelné korpusy.

Kapitola 4

Spracovanie paralelných korpusov

4.1 Kódovanie znakov

Najdôležitejšou časťou pri spracovaní paralelných korpusov je kódovanie jednotlivých textov. Najčastejšie sa v elektronických textoch používa 8-bitový kód ASCII (*American Standard Code for Information Exchange*), ktorý definuje kódy pre znaky anglickej abecedy, ale nie je vhodný pre texty obsahujúce znaky z rôznych jazykov. Preto pri práci s paralelnými korpusmi je potrebné previesť texty na vhodný typ kódovania.

Pre prácu s textami v rôznych jazykoch je možné použiť znakové sady podľa normy ISO 8859-1 až ISO 8859-15. Tie definujú šestnásť tabuliek znakov pre rôzne geografické oblasti. Napríklad ISO 8859-1 (Latin-1) je určený pre jazyky západnej Európy a ISO 8859-2 (Latin-2) pre jazyky strednej a východnej Európy. Ďalším možným riešením je použitie viacbitového kódovanie, ktoré by obsahovalo všetky možné znaky. Konkrétne sa jedná o 16-bitové kódovanie UNICODE, ktoré obsahuje viac ako 65 000 znakov. Práve toto kódovanie je pre paralelné korpusy najvhodnejšie.

Pre češtinu sa používajú v súčasnej dobe dve kódovania, ISO 8859-2 podľa medzinárodnej normy a Windows 1250 používané v operačných systémoch Microsoft. Na internete sa však stále nachádza mnoho českých textov v kódovaní ASCII obsahujúce znaky bez diakritiky, čo spôsobuje značné problémy pri zarovnávaní paralelných textov pomocou slovníka.

4.2 Formáty pre ukladanie paralelných korpusov

Teoreticky môžu mať paralelné korpusy rôznu podobu. Najčastejšie sú uložené ako súbor textov v rôznych formátoch a kódovaniach. Najjednoduchším formátom sú obyčajné textové súbory, tzv. vertikálny text, ktorý bol popísaný v práci [19]. Pre ďalšie spracovanie sa však využívajú špeciálne značkovacie jazyky ako sú SGML alebo XML charakterizované v dokumentoch [12], [14] a [13]. Mnohé paralelné korpusy sú súčasťou databázy nejakého korpusového manažéru.

4.2.1 Vertikálny text

Vertikálny text je klasický textový súbor, v ktorom sú jednotlivé elementy paralelného korpusu uložené vždy pod sebou, pričom každý element je na samostatnom riadku. Pod pojmom element paralelného korpusu rozumieme odpovedajúce si časti textov v rôznych jazykov (odstavce, vety, slová). Tie môžu byť oddelené buď znakom tabulátoru, alebo sa môžu nachádzať pod sebou v dvoch riadkoch. Štrukturálne značky sú väčšinou uvedené

v štýle SGML. Každá značka je na samostatnom riadku spolu so všetkými atribútmi, ktoré sú oddelené tabulátorom. Použitie značiek a atribútov nie je štandardizované, a preto sa jednotlivé vertikálne texty medzi sebou líšia. Ukážku krátkeho korpusu vo vertikálnom texte znázorňuje tabuľka 4.1.

Výhoda tohto formátu spočíva v jeho jednoduchosti. Je ľahko spracovateľný a dobre sa s ním manipuluje. Možno ho čítať a upravovať vo všetkých textových editoroch a prehliadačoch. Taktiež väčšina korpusových manažérov dokáže pracovať práve s týmto formátom. Veľkú podporu má aj v rôznych programovacích jazykov. Operačné systémy UNIX obsahujú mnoho štandardných programov na spracovanie takýchto typov textov (grep, cut, sed, sort, atď.).

```
<S.cz.1.1>Nemělo smysl zkoušet výtah.
<S.en.1.1>It was no use trying the lift.
<S.cz.2.1>I v lepších časech zřídka fungoval a teď se elektirický proud
cez den vypínal v rámci úsporných opatření v přípravách na Týden
nenávisti.
<S.en.2.1>Even at the best of times it was seldom working, and at
present the electric current was cut off during daylight hours.
<S.en.2.2>It was part of the economy drive in preparation for Hate Week.
<S.cz.3.1>Byt byl v sedmém patře.
<S.cz.3.2>Winston, kterému bylo devětatřicet a měl bércový vřed nad
pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul.
<S.en.3.1>The flat was seven flights up, and Winston, who was
thirty-nine and had a varicose ulcer above his right ankle, went slowly,
resting several times on the way.
```

Tabuľka 4.1: Ukážka časti korpusu vo vertikálnom texte

4.2.2 SGML

SGML (*Standard Generalized Markup Language*) je univerzálny značkovací metajazyk. Základnou textovou jednotkou v SGML dokumente sú prvky rôznych typov, ktoré sa môžu navzájom zanorovať a spolu tvoria výsledný dokument. Tieto prvky sú definované počiatočnou a koncovou značkou. Tie tvoria akoby dvojicu zátvoriek označujúcu prvok v bežnom texte. Počiatočná značka je tvorená menom prvku, ktorý sa nachádza v uhlových zátvorkách. Koncová značka má rovnakú formu s výnimkou, že za ľavou uhlovou zátvorkou sa nachádza lomítko. Ako príklad môžem uviesť `<name>`text`</name>`, kde `<name>` je počiatočná značka a `</name>` je koncová značka. Každý prvok môže obsahovať atribúty, ktorým sú priradené hodnoty. Pokiaľ chceme k nejakému prvku priradiť atribút aj s hodnotou, tak dvojicu [atribút, hodnota] vložíme do počiatočnej značky daného prvku, napríklad `<name type=1>`. Ak je prvok prázdny, môže sa použiť skrátený zápis značky `<name/>`. Príklad jednoduchého SGML dokumentu je uvedený v tabuľke 4.2.

Základnou vlastnosťou jazyka SGML je, že umožňuje deklarovať typ dokumentov. Každý dokument SGML musí odpovedať určitej formálnej špecifikácii, ktorú nazývame DTD (*Document Type Definition*). Ukážku dokumentu DTD znázorňuje tabuľka 4.3. DTD popisuje štruktúru dokumentu a zvyčajne býva uložený v oddelenom súbore. Samotný

dokument potom pozostáva z identifikácie DTD a vlastného textu doplneného o prvky. V DTD sú uvedené všetky prípustné prvky, ich prípustné atribúty a ich štruktúra (vzťahy medzi jednotlivými prvkami). Nedefinuje však konkrétny význam (sémantiku) jednotlivých prvkov, ale iba ich syntaxu. Existuje niekoľko dostupných DTD, ktoré sú štandardizované. Najznámejšou aplikáciou SGML je jazyk pre popis hypertextových dokumentov HTML (*Hypertext Markup Language*).

Medzi najviac používané štandardy pre značkovanie textu prirodzeného jazyka patrí TEI (*Text Encoding Iniciative*). Neskôr jeho úpravou vznikol ďalší štandard CES (*Corpus Encoding Standart*).

Hlavnou nevýhodou jazyka SGML je možnosť používania len obmedzeného počtu znakov (ASCII-znaky). Ostatné znaky je možné vkladať len pomocou tzv. entít, čo je postupnosť znakov začínajúcich ampersandom (&), nasleduje označenie daného znaku a bodkočiarka. To znamená, že keby sme chceli v dokumnete SGML použiť znak alphy (α), tak ho musíme zapísať pomocou entity `&alpha;`. Paralelný korpus obsahuje text v rôznych jazykoch, obsahuje teda znaky národných abecied, ktoré by sme museli zapisovať pomocou entít. Jeho výhodou je však použiteľnosť na ľubovoľnej platforme.

Pri práci so SGML dokumentom musíme ovládať daný jazyk, vedieť jeho štruktúru. Výhodou je, že existuje skupina programov, tzv. parsery, ktoré kontrolujú štruktúru SGML dokumentov na základe daného DTD. Dokážu čítať dokumenty tohto typu a vracajú jednotlivé časti textu, atribúty značiek a ich hodnoty. V súčasnej dobe existuje veľa knižníc pre prácu so SGML pre mnohé programovacie jazyky.

```
<book>
  <title>Nocna rasa< /title>
  <chapter><name>Kapitola 1 - Odhalenie< /name>
    <section>
      <sentence>Na svete existuje...< /sentence>
      <sentence>Niektory ludia...< /sentence>
      <sentence>Nie vsak...< /sentence>
    < /section>
    <section>
      <sentence>Dnes vecer...< /sentence>
      <sentence>Rozhodol som...< /sentence>
      <sentence>Bol to...< /sentence>
    < /section>
  < /chapter>
  <chapter><name>Kapitola 2 - Niet uniku< /name>
    <section>
      <sentence>Niekto zaklopal..< /sentence>
      <sentence>Hlavou mi...< /sentence>
      <sentence>Ale v tej chvili...< /sentence>
    < /section>
  < /chapter>
< /book>
```

Tabuľka 4.2: Príklad SGML dokumentu

```
<!DOCTYPE book [
  <!ELEMENT book       - -   (title?, chapter+)>
  <!ELEMENT title      - O   (#PCDATA)>
  <!ELEMENT chapter    - -   (name?, section+)>
  <!ELEMENT name       - O   (#PCDATA)>
  <!ELEMENT section    - O   (sentence+)>
  <!ELEMENT sentence   O O   (#PCDATA)>
]>
```

Tabuľka 4.3: Definícia príslušného DTD

4.2.3 XML

XML (*eXtensible Markup Language*) vznikol ako reakcia na veľkú zložitosť jazyka SGML. Je to v podstate podmnožina jazyka SGML, ale vďaka svojej jednoduchosti sa stal oveľa viac rozšírený. Umožňuje jednoduché vytváranie konkrétnych značkovacích jazykov na rôzne účely. Dokumenty môžu mať ľubovoľnú štruktúru a je v nich možné používať ľubovoľne pomenované a štruktúrované prvky. Pre definovanie štruktúry XML sa používajú jazyky pre popis schémy XML dokumentov. Tie určujú, ktoré prvky a atribúty môžeme v danom dokumente používať, ako ich môžeme vzájomne kombinovať a čo môžu obsahovať. V podstate definujeme syntaxu nového značkovacieho jazyka, ktorý má však syntaxu XML a je obmedzený na určité definované značky. Príkladom je napríklad XHTML, čo je v podstate náhrada HTML založená na XML. Pre popis schémy sa môže využiť DTD ale oveľa vhodnejšie je, keď aj schéma dokumentu je napísaná v XML. Krátka ukážka korpusu v XML dokumente je možno vidieť v tabuľke 4.4.

Dokument XML obsahuje skupinu značiek a znakov, pričom značky určujú štruktúru dokumentu a znaky predstavujú vlastný obsah dokumentu. Dokumenty XML musia spĺňať nasledujúce pravidlá:

- musí obsahovať jeden pár značiek tzv. koreňový prvok dokumentu, v ktorom sú vložené všetky ostatné prvky dokumentu
- každý prvok dokumentu musí mať počiatočnú a koncovú značku (pre prázdne elementy je možné použiť skrátený zápis)
- prvky dokumentu musia byť korektne vnorené (vnorený prvok musí byť vždy celý vo svojom nadradenom prvku, počiatočné a koncové značky jednotlivých prvkov sa nesmú prekrývať)

```
<?xml version='1.0' encoding='UTF-8'?>
<corpus>
<Cz1>Franku!</Cz1>
<En1>Hey, Frank!</En1>
<Cz2>Co takhle si pospíšit!</Cz2>
<En2>Can we hurry this up!</En2>
<Cz3>Za pět minut bude 40 pod nulou!</Cz3>
<En3>It's gonna be 40 below in five minutes!</En3>
</corpus>
```

Tabuľka 4.4: Ukážka časti korpusu v XML dokumente

4.3 Značkovanie paralelných korpusov

Iným slovom nazývané ako anotovanie paralelných korpusov. Značkovaný (anotovaný) korpus obsahuje okrem samotného textu ešte rôzne informácie, ktoré súhrnne nazývame značky. Tie môžu byť do textu pridávané ručne alebo rôznymi automatickými nástrojmi. Môžeme ich rozdeliť do štyroch skupín (kategórií): metainformácie, informácie o štruktúre textu, informácie na úrovni viet a informácie na úrovni jednotlivých slov. Bližšie sa danou tématikou zaoberá v dokumente [19].

4.3.1 Externé značkovanie (metainformácie)

Sú to všeobecné informácie o danom texte. Tieto informácie sa líšia v závislosti na typu daného korpusu. Ich význam spočíva hlavne pri výbere špecifických textov podľa zadaných kritérií. Jedná sa hlavne o bibliografické informácie, ale aj informácie o spôsobe značkovania a pod. Medzi najčastejšie používané patria:

- typ korpusu (písaný, hovorený, synchronný, diachronný, paralelný, atď.)
- názov dokumentu (či už knihy alebo článku)
- typ dokumentu (publicistický, odborný, umelecký text)
- meno autora
- popis zdroja (internet, nakladateľstvo)
- dátum publikovania
- počet viet
- a iné

Pri paralelnom korpuse nás môžu navyše zaujímať informácie, ako je jazyk daného textu, pôvodný jazyk, verzia prekladu, meno prekladateľa, atď.

4.3.2 Štrukturálne značkovanie (informácie o štruktúre textu)

Patria sem informácie o logickej a typografickej štruktúre textu. Jedná sa o hierarchické rozdelenie textu do logických celkov ako sú nadpisy, kapitoly, odstavce, vety a časti textu, ktoré nie sú riadnymi vetami, ako napríklad tabuľky. Taktiež sem patria informácie o mimoslovných prvkov vo vetách (napríklad skratky, čísla, atď.), slová cudzieho jazyka a informácie o typografickom vzhľade (riadkový zlom, stránkový zlom, rôzne typy písma). Značkovanie sa uskutočňuje pomocou štrukturálnych značiek. Tieto značky sa líšia v závislosti na danom korpuse a taktiež každá univerzita, inštitúcia má vlastné značky. Ako príklad sú nižšie uvedené niektoré značky:

- **doc** – hranice dokumentu
- **head** – nadpis
- **caption** – popisky
- **p** – odstavec
- **s** – veta, súvetie
- **language** – jazyk
- **list** – zoznam
- **table** – tabuľka
- **note** – poznámka
- **code** – úsek kódu

U niektorých značiek je vhodné si pamätať aj ďalšie informácie, ktoré sa súhrne nazývajú atribúty. Napríklad pri slovách cudzieho jazyka si môžeme pamätať v akom jazyku sú napísané, alebo pri zmene typu písma si budeme ukladať ako atribút daný typ písma.

4.3.3 Syntaktické značkovanie (informácie na úrovni viet)

Ide o značkovanie syntaktickej štruktúry. Patria sem informácie o vzájomnej závislosti medzi jednotlivými slovami a vzťahmi medzi nimi. Vznikajú syntaktické stromy (poprípade acyklické grafy) nad jednotlivými slovami. Tie vyjadrujú závislosti jednotlivých slov medzi sebou. Z každého slova buď ukazuje hrana stromu na nadradené slová alebo tvoria iba listy stromu a uzly sú akýmisi virtuálnymi uzlami, ktoré označujú čoraz väčšiu skupinu slov až nakoniec celú vetu.

Často býva označenie správnej syntaktickej štruktúry veľmi náročné, a preto sa nevyznačuje celá, ale len niektoré slovné skupiny. Potom hovoríme o čiastočnej syntaktickej analýze.

4.3.4 Morfologické a gramatické značkovanie (informácie na úrovni jednotlivých slov)

Priradenie informácií (napríklad gramatických značiek) jednotlivým slovám. Sú to slovné druhy, gramatické kategórie, atď. Jednou z najzákladnejších informácií je vyznačenie základného tvaru slova, ktoré nazývame lemma. Zvyčajne to je prvý pád jednotného čísla

pri podstatných menách, neurčitok pri slovesách, atď. Konkrétne gramatické značky sa v rôznych korpusoch líšia. Hlavným arbitrom, aké značky sa použijú je jazyk daného korpusu. Značky sú tvorené najčastejšie ako reťazec znakov (písmen a číslic), ktoré popisujú požadované gramatické kategórie. Každá inštitúcia používa vlastné značkovanie. Základnou gramatickou kategóriou je slovný druh (part of speech, POS), ktorý sa vyskytuje vo všetkých systémoch a aj jeho hodnoty sú zvyčajne u jednotlivých jazykoch rovnaké. Pre jednotlivé slovné druhy potom existujú ďalšie kategórie. Pre podstatné mená sú to rod, číslo, pád, pre prídavné mená navyše stupeň, pre slovesá osoba, rod, číslo a čas.

Okrem gramatických značiek existujú ešte negramatické značky - sémantické značky. Ide o doplnenie základných informácií a gramatických značiek, napríklad slovné vysvetlenie významu.

Nasledujúci tabuľka 4.5 zobrazuje značkovanie slov v dvoch rôznych systémoch, v „Pražskom“ (3.stĺpec) a „Brnenskom“ (4.stĺpec). Príklad bol prevzatý z dokumentu [19]. V 2.stĺpci sú uvedené základné tvary k jednotlivým slovám (lemmata).

Na	na	RR--6----------	k7c6
okně	okno	NNNS6-----A----	k1gNnSc6
seděla	sedět	VpQW---XR-AA---	k5eApFnStMmPaI
kočka	kočka	NNFS1-----A----	k1gFnSc1
,		Z:----------	
byl	být	VpYS---XR-AA---	k5eApInStMmPaI
horký	horký	AAIS1----1A----	k2eAgInSc1d1
letní	letní	AAIS1----1A----	k2eAgInSc1d1
den	den	NNIS1-----A----	k1gInSc1
,		Z:----------	
na	na	RR--6----------	k7c6
okně	okno	NNNS6-----A----	k1gNnSc6
seděla	sedět	VpQW---XR-AA---	k5eApFnStMmPaI
kočka	kočka	NNFS1-----A----	k1gFnSc1
a	a	J/----------	k8xC
koukala	koukat	VpQW---XR-AA---	k5eApFnStMmPaI
se	sebe	P7-X4----------	k3xXnSc4
ven	ven	Db----------	k6xLeAd1

Tabuľka 4.5: Priradenie značiek k jednotlivým slovám, ktoré je možné využiť v korpuse

4.4 Nástroje pre automatické značkovanie textu

Pre značkovanie textov je možné použiť dva základné postupy, a to ručné alebo automatické. Ručné značkovanie je spoľahlivé a umožňuje používať komplikovaný systém značkovania, ale na druhej strane je veľmi pracné a časovo náročné. U automatického značkovania sa nevyhneme použitiu rôznych vyspelých programov a softvérových nástrojov a následným úpravám chybne určených alebo chýbajúcich značiek. Aj napriek týmto nevýhodám sa dnes používa automatické značkovanie, pretože rozsiahle dátové súbory ako sú aj korpusy, nie je možné kvôli časovej náročnosti inak spracovať.

Dosiahnuť čo najpresnejšie označkovanie s najmenšou chybovosťou pomocou automatického značkovania je veľmi náročné a vyžaduje si to vysokú znalosť matematickej lingvistiky a to pre každý jazyk. Výber správnych morfologických a gramatických značiek z viacerých možných interpretácií slova závisí mnohokrát na konkrétnom kontexte, a okrem syntaktických faktorov tu hrajú dôležitú rolu aj sémantické. Úspešnosť automatického značkovania je ovplyvnená nie len zvolenými morfologickými nástrojmi ale aj viacznačnosťou jednotlivých slov v jazyku a úrovňou jednotlivých textov.

Existujú rôzne metódy pre značkovanie textu. Najznámejšie nástroje pre automatické značkovanie (taggers) možno rozdeliť do troch skupín podľa toho, akú metódu využívajú:

- stochastic taggers - využívajú stochastickú (štatistickú) metódu, ktorá je založená na pravdepodobnosti
- rule-bassed taggers - sú založené na syntaktických pravidlách
- hybrid taggers - tvoria kombináciu predchádzajúcich dvoch

Stochastická metóda

Je založená na pravdepodobnostiach prechodov medzi jednotlivými značkami v morfologicky analyzovanom texte. Najskôr sa ručne (správne) označkuje väčšie množstvo textu (niekoľko 1 000 slov) a vznikne tzv. tréningový korpus. Štatisticky navrhnutý značkovací program (tagger) sa „naučí“ z tréningového korpusu správne značkovanie a vytvorí si určitú predstavu o pravdepodobnostiach prechodu medzi jednotlivými značkami a ich počtom, ktorú si uloží. Na základe získaných znalostí je schopný zrealizovať vlastné označkovanie textu.

Najlepšie programy pre stochastické značkovanie korpusov v anglickom jazyku dosahujú úspešnosti okolo 97-98%. Pre český jazyk je možné dosiahnuť úroveň 94% [24]. Tento rozdiel je spôsobený v jednotlivých odlišnostiach jazykov. Anglický jazyk má pomerne pevný slovosled, a preto aj stochastické metódy založené na postupnosti značiek dosajú lepšie výsledky.

Pravidlami riadené značkovanie

Základnou podstatou je formulovanie celej rady syntaktických pravidiel, ktoré popisujú vnútorný systém daného jazyka. Pri sformulovaní nového pravidla (ktoré vzniklo z analýzy chyby) je potrebné overiť jeho správnosť na dátach korpusu. Na rozdiel od stochastickej metódy nevyžaduje tréningové dáta a ani ich nepotrebuje. Ak je nejaké pravidlo formulované so stopercentnou istotou, potom budú dáta korpusu, na ktoré sa pravidlo vzťahuje, správne označkované (pokiaľ nie je chyba v texte korpusu). Vývoj tejto metódy je v počiatočných fázach. Jej použitie sa predpokladá na jazyky s pevným slovosledom.

4.4.1 Nástroj pre značkovanie češtiny - AJKA

Ide o automatický morfologycký analyzátor [21]. Jeho hlavnou vlastnoťami je slobodná licencia (GNU/GPL), čo umožňuje jeho voľné šírenie a prístup k zdrojovým kódom. Funguje pod viacerými operačnými systémami (MS Windows, Linux). Pracuje s použitím konečného automatu. Jeho nevýhodou je nemožnosť spracovávať viacslovné výrazy (napr. Ivanka pri Dunaji).

Základný princíp ajky spočíva v rozdelení každého spracovávaného slova na 4 segmenty: prefix, koreň slova, intersegment a sufix. Medzi prefixy sú zaradené reťazce *ne-* a *naj-*, všetky ostatné prefixy sú pokladané za súčasť slova. Koreňom slova je tá časť, ktorá sa

nemení pri ohýbaní. Do intersegmentu patrí časť slova, ktorá pripúšťa alternáciu, ale nie je súčasťou žiadnej koncovky. A posledná časť slova je koncovka, ktorá sa mení pri ohýbaní.

Pri zapisovaní gramatických kategórií spracovávaných slov analyzátor ajka využíva atribútový systém, ktorý sa skladá z dvojíc znakov, kde prvý znak reprezentuje kategóriu a druhý jej hodnotu. Výhodou tohto systému je možnosť jeho rozširovania.

Kapitola 5

Zarovnanie paralelných korpusov

Pod pojmom zarovnanie (alignment) rozumieme spojenie odpovedajúcich si častí textov v dvoch alebo viacerých rôznych jazykoch. Automatické zarovnanie je veľmi často nepresné a nekompletné. Paralelné korpusy môžu byť zarovnané na úrovni viet (poprípade iných súvislých častí textu, napríklad odstavcov), alebo výnimočne na úrovni slov. Daná téma je bližšie spracovaná v dokumentoch [17] a [23]

Pri zarovnaní sa snažíme o nájdenie vzťahu 1:1, kde napríklad jedna veta v jednom jazyku odpovedá práve jednej vete v inom jazyku. V praxi však mnohokrát existujú prípady, keď jednej vete odpovedajú dve alebo viac viet, a musíme používať vzťahy typu 1:n a m:1.

5.1 Zarovnanie na úrovni viet (sentence alignment)

Pre zarovnanie viet existuje niekoľko postupov [22]. Väčšinou sa tieto postupy kombinujú, aby boli dosiahnuté čo najlepšie výsledky. Základná technika zarovnanie viet využíva špecifické znaky v textoch, ako sú napríklad interpunkčné znamienka a iné znaky (bodky, otázniky, výkričníky, čiarky, atď.). Pred samotným zarovnaním viet môže predchádzať zarovnanie na úrovni väčších celkov, ako sú napríklad kapitoly alebo odstavce.

Medzi veľmi často používané techniky patria postupy založené na porovnávaní špecifickej skupiny slov, ktoré sú rovnaké alebo podobné v oboch jazykových kontextoch. Môžu to byť mená, názvy, čísla, adresy, skratky a značky, dátumy, atď. Tieto slová sú pokladané za akési „kotevné“ slová, podľa ktorých sa určujú odpovedajúce dvojice viet. Taktiež sa môžu využívať slovníky pre preklad slov z jedného jazyku do druhého a ich následné použitie pri zarovnávaní

Ďalšie techniky sú založená na porovnávaní dĺžky viet. Základný princíp spočíva v predpoklade, že kratšie vety v jednom texte odpovedajú kratším vetám v druhom texte. Existujú však aj modely založené na počítaní znakov a slov vo vetách.

Hlavným problémom pri zarovnávaní viet je ich rozdielny počet v paralelných textoch. Veľmi často sa stáva, že jednej vete v jednom textu odpovedá viac viet v druhom texte. Hlavnou príčinou týchto nezrovnalostí je preklad súvetia na viac jednoduchých viet. Preto sa používajú rôzne techniky a ich kombinácie, aby bolo zarovnanie čo najefektívnejšie.

Zarovnanie viet je oproti zarovnaniu slov oveľa jednoduchšie a má mnohé výhody. Medzi najdôležitejšie patria menšia chybovosť a jednoduchšie spracovanie. Dokáže bez problémov nájsť odpovedajúce si vety aj v prípadoch, kde sú vety v druhom texte popísané úplne inými slovami, slovnými druhmi alebo kde je preklad formou fráz, metafor. Nevýhodou je však, že sa jedná predsa len o zarovnanie v rámci viet, a pre ďalšie spracovanie sa obvykle

musia použiť zložitejšie algoritmy a aplikácie, napríklad pre preklad alebo tvorbu slovníkov.

Vety vo výslednom korpuse sú najčastejšie očíslované a označkované v štýle SGML, napríklad `<s id=0> Veta... </s>`.

5.2 Zarovnanie na úrovni slov (word alignment)

Podstatou zarovnávania slov je rozdelenie viet na základné stavebné prvky, ako sú slová, čísla, atď. Tieto stavebné jednotky sa nazývajú tokeny a samotné rozdelenie textu tokenizácia. Pri rozdeľovaní textu do tokenov však vzniká niekoľko problémov, kde samotné rozdelenie nie je úplne jednoznačné. Napríklad český reťazec *bude-li* môže byť charakterizovaný ako jeden token, ale taktiež môže byť rozdelený na dva (*bude*, *-li*) alebo až tri tokeny (*bude*, -, *li*). Pre anglický reťazec *father's* vzniká podobný problém. Ďalšiu skupinu problémových reťazcov tvoria dátumy, skratky, adresy, atď. Dátum *1.říjen 2009* môže byť pokladaný za jeden samostatný token, ale taktiež rozdelený. Niektoré z týchto problémov sú riešené pri samotnej tokenizácii (napríklad dátumy, adresy). Iné však musia byť riešené špeciálnymi programami. Zvyčajne sa jedná o kolokácie (slovné spojenia).

Na nájdenie slovných spojení sa využívajú špeciálne algoritmy pre hľadanie MWU (multiword unit), ktoré sú popísané v dokumentoch [5] a [20]. Pre klasifikáciu spoločného výskytu slov a ich identifikácie sa používajú nasledujúce veličiny:

- **absolútna frekvencia** – udáva počet výskytov ľubovoľného slova y v zadanom kontexte slova x.

$$f(x, y)$$

- **relatívna frekvencia** – vyjadruje, koľko percent zo všetkých výskytov slova y sa nachádza v kontexte slova x.

$$f_R(x, y) = \frac{f(x, y)}{f(x)} * 100$$

- **MI-score** – vyjadruje pravdepodobnosť súčasného výskytu dvoch slov(udáva akúsi vzájomnú „príťažlivosť“), je ovplyvnený frekvenciou jednotlivých slov a preto najvyšších hodnôt dosahujú slová z nízkou frekvenciou.

$$I(x, y) = \log_2 \frac{N * f(x, y)}{f(x) * f(y)}$$

- **T-score** – miera kontrastu, popisuje vzťah medzi výskytom jednotlivých slov a ich dvojíc. Čím je T-score väčšie, tím je väčšia pravdepodobnosť, že sa jedná o ustálenú dvojicu slov (kolokáciu) a nie náhodné rozloženie.

$$T(x, y) = \frac{(f(x, y) - \frac{f(x)*f(y)}{N}}{\sqrt{f(x, y)}}$$

- **Dice-ov koeficient** - udáva pomer pravdepodobnosti výskytu dvojice slov k pravdepodobnosti výskytu iba jedného z nich. Pre slová, ktoré sú vždy spolu je koeficient 1, pre slová ktoré nikdy neboli nájdené spolu je koeficient 0.

$$Dice = \frac{2 * f(x, y)}{f(x) + f(y)}$$

$f(x)$ – frekvencia výskytu slova x
$f(y)$ – frekvencia výskytu slova y
N – veľkosť korpusu (počet slov)

Algoritmy k zarovnaniu slov používajú rôzne metódy, ktoré možno rozdeliť do nasledujúcich skupín.

5.2.1 Štatistická metóda

V súčasnosti je to najpoužívanejšie metóda. Jej hlavnou výhodou je, že je nezávislá na jazyku. Je založená na pravdepodobnosti zarovnania jednotlivých slov. Na začiatku sú slová zarovnané s nulou. Postupne sa získavajú zarovnané dvojice slov (na základe ich pravdepodobnosti). Zarovnávanie je ukončené, keď opakovaním postupu nedosiahneme už žiadne nové výsledky.

Metóda používa nasledujúce vzťahy:

- pravdepodobnosť výskytu slova x v korpuse – pomer frekvencie výskytu slova x s celkovým počtom slov v korpuse N

$$P(x) = \frac{f(x)}{N}$$

- pravdepodobnosť výskytu slova y v korpuse – pomer frekvencie výskytu slova y s celkovým počtom slov v korpuse N

$$P(y) = \frac{f(y)}{N}$$

- pravdepodobnosť, že slovo x je správna transakcia slovu y – pomer frekvencie spoločného výskytu slov x a y s celkovým počtom slov v korpuse N

$$P(x, y) = \frac{f(x, y)}{N}$$

- medzi jednotlivými pravdepodobnosťami transakcií slov platí:

$$P(x, y) = \frac{P(y, x) * P(x)}{P(y)}$$

Pred samotným zarovnávaním je použitý najskôr tzv. tréningový korpus. Pri ňom sa nastavia vhodné parametre aplikácie. Tie budú neskôr použité pri skutočnom zarovnávaní.

Medzi najčastejšie používané štatistické modely patria IBM1-6 alebo HMM word-aligment model.

5.2.2 Heuristické metódy

Metódy založené na MWU, gramatických a morfologických značkách

Pred spracovaním sú v textoch vyhľadané MWU a jednotlivým slovám sú priradené gramatické a morfologické značky. MWU sú vybraté z textu a spracované zvlášť. Slová sú zarovnávané na základe morfologických a gramatických značiek. Významnú úlohu zohráva slovný druh (POS), kde sa predpokladá, že odpovedajúce si slová sú rovnakého slovného

druhu. Nájdenie takýchto dvojíc slov môže byť ďalej využité ako „kotevné“ slová. Spojky môžu byť využité na rozdelenie viet do menších odpovedajúcich si častí.

Metódy založené na podobnosti slov

V textoch sú vyhľadané etymologicky podobné slová, reťazce - cognates, ktoré sú spárované. Podobnosť slov je založená na počte rovnakých začiatočných znakov v pomere k celkovej dlžke reťazcov.

Deklaratívny prístup

Zarovnanie prebieha v dvoch krokoch. V prvom kroku sa využívajú rôzne slovníky, prekladové databázy pre zarovnanie jednotlivých slov. Potom sú zarovnané slová odstránené z textov a prebieha druhý krok. V ňom sa môže využiť ľubovoľná iná metóda, ktorý zrealizuje zarovnanie ostávajúcich slov.

5.2.3 Ostatné metódy

Pozičná metóda

Je založená na vzájomných vzdialenostiach slov alebo pozíciách slov od začiatku vety. Predpokladá sa, že slová, ktoré stoja pri sebe v jednou jazyku majú rovnakú tendenciu aj v druhom jazyku. Najlepšie výsledky sa dosahujú v jazykoch s pevným poradím slov a príbuznými si jazykmi. Pre určovanie vzdialeností sa využívajú interpunkčné znamienka, spojky a kotevné slová.

Metóda postupného delenia

Hlavný princíp metódy je delenie odpovedajúcich si viet. Tie sa delia na menšie a menšie časti, ktoré si odpovedajú. V konečnom výsledku získame ekvivalentné slová. Pre podobné jazyky sú výsledné slová zarovnané v pomere 1:1, objavujú sa však aj zarovnania typu 1:n a m:1.

Zarovnávanie viacerých jazykov

Princíp je postavený na tranzitivite (zobrazení). Pri zarovnaní troch jazykov A, B a C môžeme predpokladať, že ak a sa zobrazí na b, a b sa zobrazí na c, tak aj a by sa malo zobraziť na c. Ak však zistíme že segment a nesúhlasí so segmentom c, tak v zarovnaní predchádzajúcich segmentoch je chyba. Metóda sa využíva pre hľadanie chybných zarovnaní.

5.3 Nástroje pre automatické zarovnávanie

5.3.1 Hunalign

Tento nástroj bol vyvinutý v rámci maďarského projektu [2] k vytvoreniu paralelného korpusu. Jeho význam spočíva v zarovnaní textov na úrovni viet. Vstupom sú dva segmentované súbory v rôznych jazykoch, kde jednotlivé segmenty sú na samostatných riadkoch. Zaujímavosťou tohto systému je, že pri zarovnávaní využíva prekladový slovník, ak však slovník nie je k dispozícii, zaobíde sa aj bez neho.

Samotné zarovnanie pozostáva z dvoch fáz. V prvej fáze je text nahrubo preložený po slovách pomocou dostupných slovníkov. Ak nie je k dispozícii slovník, pracuje sa s pôvodným textom. Následne sa vypočíta matica pravdepodobností, v ktorej sa uvažujú možné zarovnané vety len do istej vzdialenosti od diagonály matice. Odpovedajúce si vety sa

určujú pomocou pomeru dĺžok viet (algoritmus Gale-Church [10]) a počtu spoločných slov. Taktiež sú zahrnuté číselné údaje, odstavce, podobné slová, atď. Nakoniec prebehne samotné zarovnanie najpravdepodobnejšie si odpovedajúcich dvojíc na základe vstupných parametrov. V druhej fáze zarovnané dvojice typu 1:1 sú použité pre tvorbu slovníkov, ktoré sa používajú znova v prvej fáze zarovnania.

Hunalign sa nezaoberá zmenou poradia viet a nie je schopný takéto vety zarovnať. To znamená, že segmenty A a B v jednom jazyku odpovedajú segmentom B' A' v druhom jazyku.

Zhrnutie základných vlastností:

- Dokáže zarovnávať s pomocou slovníka, ale aj bez neho.
- Kombinuje zarovnanie pomocou metódy Gale-Church a na základe ekvivalentov zo slovníka.
- Umožňuje nastaviť vstupné parametre pre zarovnanie konkrétnej dvojice jazykov.
- Nedokáže vytvárať krížové korešpondencie.
- Dokáže cyklicky vylepšovať zarovnanie pomocou slovníku, ktorý si sám vytvorí.

Výstupom je súbor obsahujúci jednotlivé zarovnané segmenty alebo súbor obsahujúci len poradové čísla odpovedajúcich si segmentov.

5.3.2 GIZA++

GIZA++ je rozšírením programu GIZA, ktorý bol vyvinutý tímom Štatistického strojového prekladu v roku 1999 v centre pre spracovanie jazyka a reči na Johns-Hopkins University. GIZA++ obsahuje radu doplnených funkcií, ktoré boli navrhnuté a implementované Franz Josef Och-om.

Systém využíva pre zarovnanie slov modely IBM 1-5 a HMM word alignment model. Každý z týchto modelov je spustený pre určitý počet krokov. V každom kroku je vyhodnotené najlepšie zarovnanie slov pre každú dvojicu viet v korpuse a na základe toho sú vytvorené parametre pre nasledujúce kroky. Zarovnanie na úrovni slov je časovo náročné a to špeciálne pre veľké tréningové korpusy.

5.4 Meranie úspešnosti zarovnania

Pre meranie správnosti zarovnania (popísané v dokumentoch [8] a [17]) sa používajú nasledujúce parametre :

- **AER (Alignment Error Rata)** – percento chybne zarovnaných jednotiek.

$$AER = 1 - \frac{|A \cap P| + |A \cap S|}{|A| + |S|} * 100$$

- **Precision** - presnosť, percento určujúce počet správne určených zarovnaní z celkového počtu.

$$precision = \frac{|A \cap B|}{|A|} * 100$$

- **Recall** - pokrytie, percento správnych zarovnaní, ktoré boli rozpoznané systémom vzhľadom ku všetkým správnym zarovnaniam textu.

$$recall = \frac{|A \cap S|}{|S|} * 100$$

- **F-measure** - miera, harmonický priemer Precision a Recall.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

A – množina slov zarovnaných systémom
S – množina 100% správnych zarovnaní
P – množina pravdepodobných zarovnaní ($S > P$)

Kapitola 6

Tvorba paralelného korpusu

Na základe predchádzajúcich informácií a získaní znalostí som sa rozhodol vytvoriť menší paralelný korpus.

6.1 Zdroj dát a jeho uloženie

Pretože dosť často sťahujem titulky k filmom z rôznych serverom a k danému filmu vždy existuje niekoľko jazykových verzií a pre každý jazyk veľa prekladov, rozhodol som sa ako zdroj dát pre môj paralelný korpus využiť práve tejto možnosti. Hlavný plusom je, že zvyčajne sú titulky voľne dostupné na internete a existuje mnoho serverov, ktoré obsahujú celé databázy s titulkami. K jedným takýmto serverom s rozsiahlou databázou patrí aj server OpenSubtitles.com. Zhodou okolností jeden nemenovaný študent a kamarát nedávno získal databázu českých a anglických titulkov zo spomínaného servera, a preto som sa rozhodol využiť práve tieto titulku ako zdroj môjho paralelného korpusu a prevzal som databázu od neho.

Databáza pozostáva zo samotných titulkov, ktoré sú zabalené v archíve *tar* a uložené v dvoch adresároch. Každé titulky majú priradené identifikačné číslo (ID), podľa ktorého je pomenovaný aj archív (napríklad titulky s identifikačným číslom 1234 sa nachádzajú v archíve 1234.tar). Každé titulky patria k určitému filmu, ktorý má tiež svoje identifikačné číslo. Zoznam filmov, ich ID a zoznam titulkov k nim priradených, ktoré databáza obsahuje, sú uložené v špeciálnom súbore *export.txt*. Ku každým titulkom sú okrem ID priradené aj iné informácie, ako je jazyk titulkov, názov, dátum, atď.

6.2 Spracovanie zdrojových dát

Hlavným problémom pri práci s dátami je ich nejednoznačný formát. Titulky sú uložené v súboroch rôznych typov a každý má inú štruktúru. Preto pre ďalšie spracovanie som musel previesť titulky do jednotného formátu. Výhodou však je, že hoci sú titulky v rôznych súboroch, každý formát má danú pevnú štruktúru, takže prevod nie je komplikovaný. Väčšina titulkov je vo formátoch *sub* a *srt*, preto som sa rozhodol pracovať iba s týmito dvoma typmi súborov. Štruktúra súborov *sub* a *srt* je ukázaná v tabuľkách 6.1 a 6.2.

{1970}{2013}Měl jsem vědět,že tu budete....
{2363}{2444}Dobrý večer, profesore Dumbledore.
{2501}{2567}Tak jsou ty zvěsti pravdivé, Albusi?
{2580}{2696}Obávám se, že ano, profesorko...|...ty špatné i ty dobré.
{2696}{2745}A chlapec?

Tabuľka 6.1: Ukážka tituliek vo formáte *sub*

9
00 : 01 : 06, 712 --> 00 : 01 : 09, 837
Takto se zrodila naše rasa.

10
00 : 01 : 09, 837 --> 00 : 01 : 11, 878
Po určitou dobu jsme žili v souladu.

11
00 : 01 : 11, 920 --> 00 : 01 : 14, 670
Ale jak to bývá, velká moc
začala některé lákat...

12
00 : 01 : 14, 670 --> 00 : 01 : 18, 628
Někteří ji chtěli na stranu dobra,
jiní na stranu zla...

13
00 : 01 : 19, 778 --> 00 : 01 : 22, 778
A proto započala válka.

Tabuľka 6.2: Ukážka tituliek vo formáte *srt*

Titulky som prevádzal do jednotného formátu vždy pred ich spracovaním a pracoval som s nimi iba v rámci aplikácie. Takže išlo v podstate o štruktúru, ktorú som si vopred definoval. Konkrétne to bol zoznam komentárov, kde každý komentár obsahoval poradové číslo v rámci celých titulkov, začiatočný čas, koncový čas , priemerný čas zobrazenia komentáru a text.

Jednotlivé komentáre a k nim potrebné informácie som z titulkov získal pomocou množiny regulárnych výrazov. Časy komentárov sú uvedené v tisícinách (pri titulkoch vo formáte *srt* je potrebné jednotlivé časy prekonvertovať). Nasledujúca tabuľka 6.3 ukazuje prekonvertovanie formátu *sub* uvedeného v predchádzajúcej tabuľke 6.1 do jednotnej štruktúry.

```
[
[1],[1970],[2013],[1991,5],['Měl jsem vědět,že tu budete'],
[2],[2363],[2444],[2403,5],['Dobrý večer, profesore Dumbledore.'],
[3],[2501],[2567],[2403,5],['Tak jsou ty zvěsti pravdivé, Albusi?'],
[4],[2501],[2567],[2534],['Obávám se, že ano, profesorko...|...ty špatné
i ty dobré.'],
[5],[2696],[2745],[2720,5],['A chlapec?']
]
```

Tabuľka 6.3: Jednotná štruktúra pre spracovanie titulkov

6.3 Výber vhodnej dvojice titulkov

Po prevedení titulkov *sub* a *srt* do rovnakej štruktúry je potrebné z množiny anglických a českých titulkov pre daný film vybrať najvhodnejšiu dvojicu. Vyberie sa taká dvojica titulkov, ktoré sa najviac podobajú na základe svojho časovania. To znamená, že pre každú dvojicu českých a anglických titulkov sa nájde najprv ich posunutie v rámci časovania, a potom podľa vhodnej, vopred zadanej veľkosti časového okna sa vyberie tá najlepšia dvojica, ktorá sa najviac zhoduje.

Algoritmus na zistenie posunu časovania medzi dvojicou titulkov je postavený na metóde, ktorá hľadá posun jednotlivých špecifických znakov v rámci dvojice titulkov. Kvôli efektivite algoritmu sa hľadá posun časovania v rámci vopred definovaného intervalu. V každej z dvojice titulkov sa najprv vyhľadajú komentáre obsahujúce špecifické znaky (množinu znakov) a ich priemerný čas zobrazenia. Postupne sa pre každú dvojicu takýchto komentárov (jeden z jedných titulkov, druhý z druhých) určí ich časový posun. Nakoniec sa vyberie taký výsledný časový posun titulkov, pre ktorého interval bolo nájdených najviac časových posunov medzi komentármi. Ako špecifické znaky (skupina znakov) boli využité čiarky, otázniky, výkričníky a trojbodky. Bližšie popísaný algoritmus vysvetľuje nasledujúci príklad 1:

Príklad 1.

Tabuľky 6.4 a 6.5 predstavujú vstupné údaje. Princíp algoritmu a konkrétne riešenie pre daný príklad je predstavené v tabuľke 6.6, kde sa zisťuje posun medzi jednotlivými komentármi obsahujúcich špecifický znak '!'. Z nej je vidieť, že najviac časových posunov medzi komentármi patrí do intervalu -5500 až -4500 a preto výsledný časový posun titulkov bude približne -5000 milisekúnd.

Tabuľka 6.4: Časť komentárov z anglických titulkov spolu s priemerným časom zobrazenia.

```
EN1   12300   I never touched the money.
EN2   13400   FBI!
EN3   16800   You're under arrest.
EN4   17900   I'm Rainbow-Randolph!
```

Tabuľka 6.5: Časť komentárov z českých titulkov spolu s priemerným časom zobrazenia.

CZ1	16500	Pusťte kufr.
CZ2	17400	Ani jsem se těch peněz nedotkl.
CZ3	18550	FBI!
CZ4	21900	Jste zatčen.
CZ5	23070	Já jsem Duhovej Randolph!

Tabuľka 6.6: Ukážka princípu algoritmu pre hľadanie posunu medzi titulkami.

$$\begin{aligned}
\texttt{EN2}-\texttt{CZ3} &= 13400 - 18550 = -5150 \\
\texttt{EN2}-\texttt{CZ5} &= 13400 - 23470 = -10070 \\
\texttt{EN4}-\texttt{CZ3} &= 17900 - 18550 = -650 \\
\texttt{EN4}-\texttt{CZ5} &= 17900 - 23070 = -5170
\end{aligned}$$

Po určení časového posunu titulkov treba zo všetkých dvojíc vybrať tie najvhodnejšie. Riešil som to jednoduchým algoritmom, ktorý sa snažil priradiť postupne komentáru v jedných titulkoch komentár v druhých titulkoch na základe ich časovania, posunu a časového okna. Dvojica titulkov, v ktorých bolo k sebe priradených čo najviac komentárov bola určená ako najvhodnejšia a ďalej spracovávaná.

6.4 Čistenie titulkov

Niektoré formáty titulkov pripúšťajú aj značky v texte, ktoré menia formát a farbu písma pri ich spracovaní rôznymi prehrávačmi. Preto pred samotným zarovnaním som sa snažil odstrániť tieto značky a iné neželané znaky, napríklad `|`, `<i> ... </i>`, atď. Ich odstránenie som riešil pomocou regulárnych výrazov.

6.5 Hrubé zarovnanie

Ide o zarovnanie komentárov pomocou ich časovania bez použitia slovníkov. Postupne pre každú dvojicu titulkov priraďuje k sebe anglické a české komentáre na základe ich priemerného času zobrazenia, časového okna a posunu titulkov.

Výsledkom hrubého zarovnania je paralelný korpus, ktorý obsahuje zarovnané titulky na úrovni komentárov v pomere 1:1, 1:m, n:1 a m:n. Pretože sa jedná o zarovnanie iba na základe časovania, môže sa stať, že niektorým komentárom z jedného jazyka budú priradené nesprávne komentáre z druhého jazyka, ktoré mali byť priradené predchádzajúcemu alebo nasledujúcemu komentáru.

Výhodou tohto typu zarovnania je jeho nezávislosť na jazyku a môže byť použitý aj pre iné jazyky ako len češtinu a angličtinu. Vo výsledku sa však musíme zmieriť s jeho nedostatkami.

6.6 Jemné zarovnanie

Zarovnanie používajúce slovník k odstráneniu nepresností, ktoré vznikli v predchádzajúcom kroku (v hrubom zarovnaní). Konkrétne sa jedná o odstránenie väčšiny vzťahov typu m:n a čiastočnú kontrolu spárovaných komentárov. Pri hrubom zarovnaní môže dôjsť k spárovaniu takej dvojici komentárov, kde ku komentáru v jednom jazyku bol priradený nesprávny komentár z druhého jazyku, ktorý je predchodcom alebo nasledovníkom správneho. Nasledujúci problém je ukázaný v tabuľke 6.7. Jemné zarovnanie sa snaží o opravenie takýchto chybných dvojíc.

```
<eng comentary=332>You remember his name?
<cze comentary=321>Pamatujete si to jméno?  Luis Uribe.
<eng comentary=333>Luis Uribe.  Why does he hate Americans?
<cze comentary=322>Proč nenávidí Američany?
```

Tabuľka 6.7: Ukážka nesprávnej korešpondencii pri hrubom zarovnaní.

Pred samotným jemným zarovnaním je upravený vstup (vstupom pre jemné zarovnanie je výstup hrubého). Postupne sú rozdelené všetky komentáre na vety. Inak povedané, ak niektorý komentár zo zarovnanej dvojice pozostáva z dvoch alebo viacerých viet, tak sa musí rozdeliť. V nasledujúcej tabuľke 6.8 je uvedené rozdelenie komentárov z tabuľky 6.7 na vety.

```
<eng comentary=332>You remember his name?
<cze comentary=321>Pamatujete si to jméno?
<eng comentary=>
<cze comentary=321>Luis Uribe.
<eng comentary=333>Luis Uribe.
<eng comentary=>
<eng comentary=333>Why does he hate Americans?
<cze comentary=322>Proč nenávidí Američany?
```

Tabuľka 6.8: Ukážka rozdelenia komentárov na vety.

Následne je potrebná tokenizácia. Ide o rozdelenie viet na jednotlivé tokeny. Pre toto rozdelenie som využil modul *regexp*, ktorý sa nachádza v balíčku *nltk.tokenize* [15]. Ten je súčasťou NLTK (*Natural Language Toolkit*) a je voľne dostupný na internete. Za tokeny som pokladal jednotlivé slová, celé čísla a desatinné čísla. Pre získanie základného tvaru českých slov som využil morfologický analyzátor LIBMA dostupný na FIT VUT v Brne. Pre anglické slová som použil modul *morphy*, ktorý je súčasťou balíčka *nltk.corpus.reader.wordnet* [16]. Ak pre daný token bolo nájdených viacej základných tvarov, tak pracujem s celou ich množinou, ak nebol nájdený žiadny, pracujem s tokenom v pôvodnej podobe. Postupne získam všetky základné tvary slov pre jednotlivé tokeny vo vetách.

Samotné jemné zarovnanie spočíva v postupnom prechádzaní spárovaných komentárov (ktoré sú už rozdelené na vety) a v ich kontrole. Pre každú anglickú a odpovedajúcu českú vetu sa zistí pomocou slovníku počet slov (resp. ich základných tvarov), ktoré si odpovedajú. Tento počet sa zisťuje nie len pre danú dvojicu spárovaných viet, ale aj pre susedné

vety v danej dvojici (t.z. že počet sa zisťuje medzi anglickou a priradenou českou vetou, a ešte aj medzi danou anglickou vetou a predchádzajúcou a nasledujúcou vetou pridelenej českej vety). Dochádza k zarovnaniu tej dvojice viet, kde bol počet odpovedajúcich si slov najväčší. Postupne sa prekontrolujú a zarovnajú všetky dvojice viet, ktoré boli získané úpravou výstupu hrubého zarovnania.

Výstup jemného zarovnania sa ukladá pre každú dvojicu titulkov zvlášť. Časť výstupu je znázornená v tabuľke 6.9. Výsledný zarovnaný korpus obsahuje nasledujúce značky:

- **metainformácie**
  <`film ID=`n > – ID-číslo filmu, ku ktorému patria zarovnané titulky
  <`subtitles ID=`n `language=cz count_talk=`m > – ID-číslo českých titulkov a počet obsahujúcich tokenov
  <`subtitles ID=`n `language=eng count_talk=`m > – ID-číslo anglických titulkov a počet obsahujúcich tokenov

- **informácie na úrovni viet**
  <`eng comentary=`n > – poradové číslo anglického komentáru v titulkoch, z ktorého pochádza daná veta
  <`cze comentary=`n > – poradové číslo českého komentáru v titulkoch, z ktorého pochádza daná veta

Pri jemnom zarovnaní sa taktiež generuje zoznam slov (spolu s vetami kde sa nachádzajú), ktoré neboli nájdené ani v morfologických slovníkoch ani v normálnom dvojjazyčnom slovníku. Časť výstupu je možné vidieť v tabuľke 6.10.

```
<film ID=10>
<subtitles ID=12630 language=cze count_talk=8652>
<subtitles ID=151871 language=eng count_talk=10219>

<eng comentary=1>It‘s starting!
<cze comentary=2>Už to začíná!

<eng comentary=1>Rainbow Randolph!
<cze comentary=2>Duhový Randolph!

<eng comentary=1>Welcome to The Rainbow Randolph Show!
<cze comentary=2>Vítejte do show Duhového Randolpha!

...
</film>
```

Tabuľka 6.9: Ukážka výstupu jemného zarovnania.

```
<film ID=10>
<subtitles ID=12630>

<comentary=2>randolph    Duhový Randolph!
<comentary=3>randolpha   Vítejte do show Duhového Randolpha!

...
< /film>
```

Tabuľka 6.10: Ukážka výstupu vygenerovaného zoznamu slov.

Kapitola 7

Výsledky a štatistický pohľad na korpus

Z celkového počtu boli spracované titulky k 3 000 rôznym filmom. Nie však ku každému filmu boli nájdené titulky v oboch jazykoch (českom i anglickom) a k niektorým filmom sa taktiež nepodarilo nájsť vhodnú dvojicu tiulkov na spracovanie, preto je aj počet spracovaných dvojíc nižší. Celkovo sa podarilo zarovnať niečo len cez 50%, konkrétne 1569 dvojíc. Tento počet sa však ešte zmenšil po ručnej kontrole, kde boli odstránené zle zarovnané titulky (veľkú časť však tvorili také dvojice, ktoré obsahovali iný jazyk ako bolo predpokladané vzhľadom na chybu v databáze). V celkovom výsledku boli správne zarovnané dvojice titulkov k 1379 filmom. Spolu obsahujú približne 7 019 822 tokenov v českých a 8 859 447 tokenov v anglických titulkoch.

V priemere 95% zo zarovnaných dvojíc viet je vzťahu 1:1. Ostatné zarovnané dvojice sú vzťahu 1:m, n:1 a príležitostne m:n. Výsledná kvalita zarovnanie bola ohodnotená podľa vzťahu precission, ktorej hodnota je približne 98%. To znamená, že zo všetkých zarovnaných dvojíc je približne 98% správne zarovnaných.

V celom procese zarovnania bolo spolu vygenerovaných 432 186 českých a 339 311 anglických slov, ktoré neboli nájdené v slovníkoch morfologických analyzátorov a ani v dvojjazyčných slovníkoch. Väčšina z týchto slov sú buď názvy, mená alebo preklepy.

V zrovnaní s paralelným korpusom OpenSubtitles [4] je môj výsledný korpus len malé zrnko. Spomínaný korpus zahŕňa 30 jazykov a spolu obsahuje približne 20 400 zarovnaných dvojíc titulkov, ktoré sú uložené aj vo vertikálnom texte ale aj ako XML dokumenty. Celkovo sú tvorené 149 436 587 tokenmi. Najväčšie zastúpenie z jazykov majú angličtina, španielčina, portugalčina a čeština, najmenšie zasa lotyština, litovčina a japončina. Pre zrovnanie s mojím korpusom, obsahuje 19 833 493 anglických a 11 221 227 českých tokenov.

Kapitola 8

Ďalšie práce

8.1 Rozšírenie korpusu

Po dokončení praktickej časti mojej práce ma napadlo mnoho vylepšení a možných rozšírení, ktoré by určite pridali na jej kvalite. Ako prvé ma napadlo rozšírenie databázy titulkov nie len na dvojicu jazykov angličtina-čeština, ale aj iné jazyky a vytvárať tak multijazyčný paralelný korpus, kde by boli navzájom priradené časti textov z viacerých jazykov.

Druhým zaujímavým rozšírením projektu je automatické získavanie dát z rôznych serverov obsahujúcich celé databázy titulkov. Túto metódu som z počiatku zavrhol kvôli obmedzeniam na strane servera, kde bolo možné za určitý čas stiahnuť iba obmedzené množstvo titulkov. Ale som toho názoru, že po určitej dohode z rôznymi servermi by bolo možné uskutočňovať občasné aktualizácie a sťahovať nové titulky a rozširovať tak výsledný korpus.

Pri spracovaní projektu mi hlavou preletela otázka, či by bolo možné použiť dabingy k filmom k vytvoreniu paralelného hovoreného korpusu, ktorý by bol zarovnaný na jednotlivé krátke úseky filmu. Zaujímavé by však tiež bolo sa pokúsiť zarovnať titulky s dabingom a vytvárať tak zvukové podoby jednotlivých slov. To je ale už iná kapitola, kde zatiaľ moje vedomosti nesiahajú a ani neviem, či niečo podobné sa dá zrealizovať.

8.2 Ďalšie využitie a zlepšenie

Jednou z možných úprav korpusu je určite rozšírenie značiek na úrovni jednotlivých slov. To by prinášalo možnosť generovania nových slovníkov. Pri prezeraní výsledkov z mojej aplikácie, konkrétne vygenerovaných slov, ktoré neobsahoval morfologický analyzátor ani slovník, som zistil, že väčšina nových slov sú buď mená alebo preklepy v textoch. To by sa dalo využiť pri generovaní nejakej databázy mien alebo pri oprave gramatických chýb v jednotlivých titulkoch.

Ďalšou možnosťou využitia by som videl v automatickom preklade titulkov z jedného jazyka do druhého. To by však vyžadovalo komplexnejšie znalosti v danej oblasti a taktiež dobrú znalosť daných jazykov.

Kapitola 9

Záver

Cieľom mojej práce bolo vytvoriť jednoduchú aplikáciu, ktorá by bola schopná spracovávať databázu titulkov a vytvárať jednoduchý paralelný korpus. Domnievam sa, že požadovaný úkol som splnil, i keď sa dá v mojom diele určite ešte veľa vecí vylepšovať. Hlavnú úlohu zohral určite aj čas, ktorý mi neumožnil dotiahnuť prácu do „dokonalosti“. Pri opakovanom spracovaní projektu by som určite postupoval zasa trošku inak, ale základné princípy by boli pravdepodobne rovnaké. V priebehu práce som niekoľkokrát musel prepisovať často celé algoritmy kvôli veľkej časovej náročnosti, ale s konečným výsledkom som osobne spokojný.

Pri tvorbe korpusu som sa naučil mnoho nových vecí a uvedomil si zložitosť jednotlivých jazykov. Hoci ich denne používam, či už v písanej podobe alebo v bežnej komunikácii, nikdy som si neuvedomil, ako komplikované môže byť ich spracovanie a analýza jednotlivých slov. Korpus, ktorý som vytvoril má mnohé využitie. Rád by som ho sám niekedy v budúcnosti použil pre hlbšiu analýzu jednotlivých jazykov.

Pri spracovaní projektu som si sám stanovil cieľ, aby výsledná aplikácia bola použitá nie len na dvojicu jazykov angličtina a čeština, ale na ľubovoľnú dvojicu jazykov. Tento cieľ sa mi z časti podarilo splniť, hoci nie je možné použiť morfologickú analýzu slov a slovníky pre iné jazyky, ale aj bez týchto častí je možné dosiahnuť pozoruhodné výsledky a vytvoriť celkom slušný paralelný korpus.

Literatúra

[1] *Všeobená encyklopédia Diderot*. Praha: Diderot, 1999.

[2] Hunalign - sentence level aligner [online]. `http://mokk.bme.hu/resources/hunalign`, [cit. 2009-5-10].

[3] Kačenka [online]. `http://www.phil.muni.cz/angl/kacenka/kachna.html`, [cit. 2009-5-10].

[4] OpenSubtitles [online]. `http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php`, [cit. 2009-5-10].

[5] Český národní korpus: Paralelní korpusy [online]. `http://ucnk.ff.cuni.cz/bonito/stat.php`, [cit. 2009-5-10].

[6] Wikipedia – The Free Encyclopedia: Text corpus [online]. `http://en.wikipedia.org/wiki/Text_corpus`, [rev. 2009-4-9][cit. 2009-5-10].

[7] Bournard, L.: About the British National Corpus [online]. `http://www.natcorp.ox.ac.uk/corpus/index.xml?style=.pdf`, január 2009 [rev. 2009-2-3][cit. 2009-5-10].

[8] Brockett, C.; Dolan, W. B.: Support Vector Machines for Paraphrase Identification and Corpus Construction [online]. `http://www.aclweb.org/anthology-new/I/I05/I05-5001.pdf`, [cit. 2009-5-10].

[9] Francis, W.; Kučera, H.: Brown Corpus Manual [online]. `http://icame.uib.no/brown/bcm.html`, 1964 [cit. 2009-5-10].

[10] Gale, W. A.; Church, K. W.: A Program for Aligning Sentences in Bilingual Corpora [online]. `http://acl.ldc.upenn.edu/J/J93/J93-1004.pdf`, [cit. 2009-5-10].

[11] Germann, U.: Aligned Hansards of the 36th Parliament of Canada [online]. `http://www.isi.edu/natural-language/download/hansard/`, 2001 [cit. 2009-5-10].

[12] Hruška, T.; Burget, R.: Internetové aplikace (WAP) II. – část SGML, HTML, CSS, DOM [online]. `https://www.fit.vutbr.cz/study/courses/WAP/private/opory/OporaWAP2SGMLHTMLCSSDOM.pdf`, 2007-2-15 [cit. 2009-5-10].

[13] Hruška, T.; Burget, R.: Internetové aplikace (WAP) II. – část XML, XML schémata, XPath, XSLT [online]. `https://www.fit.vutbr.cz/study/courses/WAP/private/opory/OporaWAP3XMLXPathXQueryXSLT.pdf`, 2007-3-18 [cit. 2009-5-10].

[14] Kosek, J.: SGML: Standard Generalized Markup Language [online]. http://www.kosek.cz/clanky/cw/sgml.html, 1999 [cit. 2009-5-10].

[15] NLTK: Modul regexp [online]. http://docs.huihoo.com/nltk/0.9.5/api/nltk.tokenize.regexp-module.html, [cit. 2009-5-10].

[16] NLTK: Type WordNetCorpusReader [online]. http://nltk.googlecode.com/svn/trunk/doc/api/nltk.corpus.reader.wordnet.WordNetCorpusReader-class.html, [cit. 2009-5-10].

[17] Rosen, A.: Paralelní korpusy [online]. http://utkl.ff.cuni.cz/~rosen/public/parcorp_print.pdf, 2006 [cit. 2009-5-10].

[18] Rosen, A.: Paralelní korpus Intercorp [online]. https://trnka.ff.cuni.cz/ucnk/intercorp/?req=id:9, [rev. 2009-5-5] [cit. 2009-5-10].

[19] Rychlý, P.: *Korpusové manažery a jejich efektivní implementace*. Dizertačná práca, Masarykova univerzita v Brně – Fakulta informatiky, únor 2002.

[20] Rychlý, P.: A Lexicographer-Friendly Association Score [online]. http://nlp.fi.muni.cz/raslan/2008/papers/13.pdf, [cit. 2009-5-10].

[21] Sedláček, R.: *Morfologický analyzátor češtiny*. Diplomová práce, Masarykova univerzita v Brně – Fakulta informatiky, 1999.

[22] Veronis, J.: *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer Academic Publishers, iSBN 0792365461.

[23] Čechová, H.: *Korpusová lingvistika – Problematika česko-italského paralelního korpusu*. Magisterská diplomová práce, Masarykova univerzita v Brně, 2006.

[24] Čermák, F.; Schmiedtová, V.: Český národní korpus – Základní charakteristika a širší souvislosti [online]. http://knihovna.nkp.cz/pdf/0403/0403152.pdf, 2004 [cit. 2009-5-10].

Dodatok A

Obsah CD

Na CD sa nachádzajú nasledujúce prílohy:

- bakalarka.pdf – bakalárska práca vo formáte pdf
- subtitles.py – vlastná knižnica obsahujúca rôzne funkcie pre prácu s titulkami
- skript.py – skript určený pre zarovnanie titulkov
- configuration.py – konfiguračný súbor aplikácie
- README – popis aplikácie a ďalších súborov nachádzajúcich sa na CD
- plagat/ – adresár obsahujúci plagát k bakalárskej práci v rôznych formátoch
- output/ – adresár obsahujúci výsledný korpus