Predstavitev tem za seminarske naloge
17.3.2006

Kaj so seminarske naloge

Seminarska naloga je obvezen izdelek, ki ga študenti izdelajo samostojno ali v skupini v skladu z navodili spodaj in ga ustno predstavijo v razredu. Seminarska naloga se ocenjuje in sestavlja 25% končne ocene pri predmetu.

Seminarska naloga je sestavljena iz praktičnega dela in poročila. Pisno poročilo naj se drži običajne oblike za znanstvene članke, vsebuje naj torej naslov, avtorje, povzetek, uvod (ozadje problema), poročilo o izvedbi, zaključke, literaturo in (če so) priloge. Okvirni obseg je 8 strani za individualne seminarske naloge, več za skupinske.

Predlagane teme

Spodaj je naštetih nekaj predlogov za teme, ki vključujejo osrednje jezikovnotehnološke probleme in se deloma navezujejo na tekoče projekte, predstavljene pri predavanjih. Študentje se tako spoznajo s sodobnimi pristopi k računalniški obravnavi slovenščine, njihovi rezultati pa bodo lahko tudi koristni. Nekatere predlagane naloge so razmeroma kompleksne, zato jih bo laže reševati v skupini. Možno je tudi več individualnih nalog na isto temo, vendar z različnimi podatki, dobrodošli pa so tudi individualni predlogi tem.

1. tema
Oblikoslovno označevanje besedila

V okviru naloge je potrebno pregledati in popraviti napake v avtomatsko oblikoslovno označenem slovenskem besedilu ter napake analizirati. Označevanje je tako kot v korpusu MULTEXT-East, vsaka beseda v besedilu je torej označena z lemo in oblikoslovno oznako. Podatki bodo na voljo v formatu Excel, zato bo potrebno poznavanje delovanja te razpredelnice.

Nekaj možnih besedil:
- 4.000 besed iz pravnega reda EU
- 3.000 besed iz časopisnih člankov
- 2.000 besed iz knjižnice AHlib

Literatura:
- (predavanje o označevanju 7.4.2006)
- spletne strani MULTEXT-East, http://nl.ijs.si/ME/ in oblikoslovne specifikacije, http://nl.ijs.si/ME/V3/msd/
- Tomaž Erjavec in Sašo Džeroski (2004) Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. Applied Artificial Intelligence 18(1), pp. 17-40.
- Birte Lönneker (2005): Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo. Slavistična revija 2/2005. 193-210.

2. tema
Korekcije besedil slovenske literature iz XIX stoletja

V okviru naloge je potrebna korigirati in oblikovati besedila za knjižnico AHLib, zahteva pa obvladovanje urejevalnika Word in razpredelnice Excel. Korekcije se izvajajo v urejevalniku Word, nato pa se jih preko mrežnega servisa pretvori v standardni zapis v XML/TEI, iz njega pa v HTML. V besedilu se popravlja napake nastale v procesu OCR, označuje strukturo besedila: strani, poglavja, tujejezični citati, itd. Za kontrolo dela služi avtomatska pretvorba v HTML. Na mrežnem servisu se besedila tudi jezikoslovno obdelajo, vse neznane besede pa se, skupaj z oznakami izpišejo kot konkordance v razpredelnici Excel. Tu je potrebno popraviti nepravilno lematizirane besede in lemam dopisat sodobno ustreznico.

Literatura:
- (predavanje o standardih za zapis korpusov 14.4.2006)
- Tomaž Erjavec, Matija OGRIN (2005). *Digitalisation of literary heritage using open standards*. In: Innovation and knowledge economy: issues, applications, case studies, (Information and communication technologies and the knowledge economy). Amsterdam [etc.]: IOS Press, 2005, str. 999-1006.
- Tomaž Erjavec and Sašo Džeroski (2004) Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence 18(1)*, pp. 17-40.

3. tema
Skladenjsko označevanje

Z odvisnostnimi drevesi je potrebno označiti slovensko besedilo. Kot model se uporablja Prague Dependency Treebank, kot orodje pa urejevalnik TrEd (potrebna je instalacija tega programa.). Kot vir primerov se uporabi do sedaj narejeni korpus SDT.

Možna besedila:
- 50 stavkov »1984«
- 70 stavkov časopisnih člankov
- 100 stavkov pravnega red EU

Literatura:
- Predavanje o skladnji 19.5.2006
- Bajič J., Panevová, J., Buránová, E., Urešová, Z., Bémová, A. (1997) *A Manual for Analytic Layer Tagging of the Prague Dependency Treebank*. UFAL Technical Report TR-1997-03, Charles University, Czech Republic.
- Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky, Andreja Žele (2006). *Towards a Slovene Dependency Treebank*. In Proceedings of the Conference on Language Resources and Evaluation, Paris, ELRA.

4. tema
Korpusna analiza literarnega dela ali slogovna primerjava dveh slovenskih avtorjev

Cilj naloge je, s pomočjo Wordsmitha in morebitnih drugih orodij analizirati jezikovne značilnosti izbranega literarnega dela in predstaviti slogovne posebnosti določenega avtorja.

Besedila so na voljo na spletni strani http://www.ijs.si/lit/leposl.html-l2. Tu si izberete eno ali več del določenega avtorja. Pri shranjevanju besedila na računalnik se vam utegnejo šumniki spremeniti v karkoliže, zato besedilo pred uporabo z WSmithom ustrezno popravite in shranite v Windows kodiranju.

V uvodnem delu naloge se spodobi napisati kak znak o avtorju, denimo letnice rojstva in v katero obdobje slovenske književnosti sodi.

Med možne elemente analize sodijo:

1. Osnovna besedilna **statistika**, razmerje med pojavnicami in različnicami, dolžina stavkov itd. Če želite o tem povedati kaj konkretnega, se splača vaše besedilo primerjati s kakim drugim besedilom, morda tudi z več drugimi avtorji.
2. **Pogoste in redke besede**. Med pogostimi se bodo morda našle ključne osebe knjige ali še kaj drugega, kar razkriva slog avtorja, med redkimi pa kake besede, ki jih morda uporablja le ta avtor. Izmed pogostih besed se splača izbrati nekaj kandidatov, ki bi bili primerni za podrobnejši pregled (sumimo, da bodo kake zanimive kolokacije ali fraze).
3. Večbesedni **skupki in kolokacije** okrog besed, ki ste jih v prejšnji točki izbrali za kandidate.
4. Ključne besede (**keywords**). Zelo luštna funkcija WSmitha, pri kateri nam na površje splavajo besede, tipične za neko besedilo.
5. Spreminjanje sloga pri različnih delih istega avtorja. S primerjavo osnovne statistike ugotovite razlike v gostoti besedišča, dolžini stavkov in besed itd.
6. Spreminjanje sloga v enem literarnem delu. Delo razsekate na več kosov in naredite besedilno statistiko za vsak kos posebej. Primerjava kaže spreminjanje sloga.
7. Razporeditev slogovno zanimivih elementov po besedilu (s funkcijo **Plot**). Slogovno zanimivi elementi so lahko tudi denimo medmeti, vzkliki, čustveno obarvani izrazi, slengizmi, itd., skratka karkoli, kar je značilno za avtorja.
8. Primerjava vseh zgornjih elementov pri dveh izbranih avtorjih.

5. tema
Analiza razlik med izvirnimi in prevedenimi besedili

Cilj naloge je korpusno podprta analiza razlik med izvirnimi in prevedenimi besedili v slovenskem jeziku. Naloga obsega izdelavo korpusa izvirnih in prevedenih besedil ter njuna analiza s pomočjo programa Wordsmith. Možni elementi analize obsegajo:

1. Razmerje med pojavnicami in različnicami (TTR), in analiza leksikalne gostote,
2. analiza frekvenčnega besednega seznama in primerjava relativnih pogostosti,
3. prisotnost pojasnjevalnih in razlagalnih struktur v prevedenih besedilih,
4. analiza enopojavnic.

Literatura:

Mona Baker: A corpus-based view of similarity and difference in translation. International Journal of Corpus Linguistics, Volume 9, Number 2, 2004, pp. 167-193(27).

6. tema
Jezikoslovna analiza jezikovnih sprememb v obdobju od XX do XX (dve primerni obdobji glede na besedila, ki so na voljo)

Cilj naloge je, s pomočjo orodja Wordsmith analizirati leksikalne, skladenjske, slogovne in besediloslovne spremembe v določenem časovnem obdobju. Naloga obsega izdelavo korpusa, ki ga je mogoče sestaviti iz besedil slovenskega leposlovja (http://www.ijs.si/lit/leposl.html-l2) ali drugih elektronsko dostopnih besedil (Ahlib itd., v dogovoru s predavatelji), nato pa diahrono analizo vseh omenjenih ravni. Za ugotavljanje razvojnih tendenc bo korpus treba najprej razdeliti na manjše dele po kronoloških obdobjih, nato analizirati vsak del posebej, in nazadnje primerjati dele med seboj ter opisati razvojne težnje.

Med možne elemente analize sodijo:

- Osnovna besedilna **statistika**, razmerje med pojavnicami in različnicami, dolžina stavkov itd.
- **Pogoste in redke besede**. Izmed pogostih besed se splača izbrati nekaj kandidatov, ki bi bili primerni za podrobnejši pregled (sumimo, da bodo kake zanimive kolokacije ali fraze).
- Večbesedni **skupki in kolokacije** okrog besed, ki ste jih v prejšnji točki izbrali za kandidate.
- Ključne besede (**keywords**). Zelo luštna funkcija WSmitha, pri kateri nam na površje splavajo besede, tipične za neko besedilo.
- Spreminjanje sloga skozi čas. Korpus razsekate na več kosov in naredite besedilno statistiko za vsak kos posebej. Primerjava kaže spreminjanje sloga.
- Razporeditev slogovno zanimivih elementov po besedilu oz. korpusu (s funkcijo **Plot**). Slogovno zanimivi elementi so lahko tudi denimo medmeti, vzkliki, čustveno obarvani izrazi, slengizmi, itd.

7. tema
Učenje jezikov in jezikovne tehnologije

Ta tema zajema manj praktičnega dela in več pregledovanja spletnih virov in literature na to temo. Cilj je izdelati pregledno študijo, kje, kako in zakaj se učenje jezikov uspešno povezuje z jezikovnimi tehnologijami. Pri tem je smiselno najprej pregledati tehnologije, ki se uporabljajo za učenje angleščine kot tujega jezika, ker je takih aplikacij verjetno največ, nato pa raziskavo razširiti na druge jezike, tudi na slovenščino. Pisni del naloge naj bi obsegal med 10 in 15 strani pregleda, v predstavitvi pa naj bi kolegom predstavili najpomembnejše spletno dostopne aplikacije za učenje jezikov.

Literatura ni posebej določena, viri se večinoma najdejo na spletu.

8. tema
Korpusno zasnovana leksikalna analiza

Cilj naloge je podrobno semantično analizirati par slovenskih besed. S pomočjo korpusov FIDA, FIDAplus in Nova Beseda je potrebno identificirati njihove različne pomene, te rezultate pa primerjati z opisom gesel v SSKJ-u, ter predlagati spremembe. Potrebno je biti pozoren tako na slovnične lastnosti besed kot njihove kolokacije.
Za analizo lahko izberete:

- par navidezno sinonimnih besed ter analizirate razlike in podobnosti v njihovih pomenih in rabah
- več besed, ki so izpeljane iz enega zadosti pogostega korena
- vkolikor je analiza zelo podrobna in natančna, je lahko usmerite na eno samo besedo.

Literatura:

- Hanks, Patrick (2002). Mapping Meaning onto Use. Marie-Helene Correard (ed): Lexicography and Natural Language Processing: A Festshrift in Honour of B.T.S Atkins, str.156-198. United Kingdom, EURALEX.
- Gorjanc, Vojko, Krek, Simon in Gantar, Apolonija (2005). Slovenska leksikalna podatkovna zbirka. Jezik in slovstvo, 50/2. 3-20.
- Hanks, Patrick (1997). Lexical sets: relevance and probability. Translation and Meaning. Part 4. School of Translation and Interpreting, Maastricht, The Netherlands, tudi dostopno na www.patrickhanks.com/papers/LexicalSets.pdf
- Moon, Rosamund (1987). The Analysis of Meaning. Sinclair J.M., urednik. Looking up: An Account of the COBUILD Project in Lexical Computing. Collins, London. 86–115

9. tema
Primerjava obstoječih korpusov za slovenski jezik

Cilj naloge je primerjava pokritja in zasnove mrežnih konkordančnikov slovenskega jezika. Naloga je pregledna, pa tudi praktična, saj predvideva izbiro nekoliko slovenskih besed (npr. 2 samostalnika, 2 glagola, 2 pridevnika) in analizo njihovih pojavitev v korpusih FIDA, FIDAplus, Nova Beseda, ter tistih, dostopnih na IJS. Predstavite razlike v načinih iskanja, ki jih ponujajo korpusi, ter razlike v dobljenih rezultatih.

Literatura:

- Beseda: http://bos.zrc-sazu.si/
- FIDA: http://www.fida.net/
- FIDA+: http://www.fidaplus.net/
- nl@ijs: http://nl2.ijs.si/
- Gorjanc, Vojko (2005). Uvod v korpusno jezikoslovje. Domžale, Izolit.

10. tema
Izdelava spletnega slovarčka v TEI

Izbrati je potrebno seznam besed (20-50 besed) ter za njih narediti slovarček, ki obsega npr. definicijo, primere iz korpusa, prevode v druge jezike. Slovar zapišite v XML, po standardu TEI. Za urejevalnik se lahko uporabi kar Wordpad, ali pa demo verzija (deluje en mesec) katerega od komercialnih urejevalnikov XML. Za preverjanje se lahko uporabi katerega od brskalnikov, ali pa kak drug validator.

Literatura:

- Spletna stran TEI http://www.tei-c.org/
- Erjavec, Tomaž, Hmeljak Sangawa, Kristina, Srdanović, Irena, Vahčič, Anton Ml. Making an XML-based Japanese-Slovene Learners' Dictionary. In Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC'04, 26-28 May 2004. Lisbon.
- Krek, Simon (2003). Sodobna dvojezična leksikografija. Jezik in slovstvo, 48/1. 45-60.