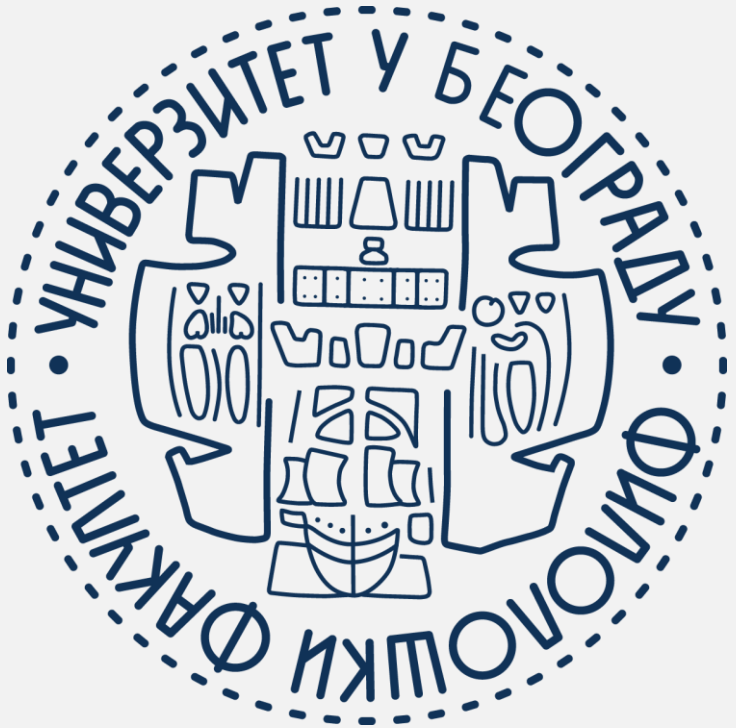
Anotacija „ITALSERB“ korpusa

Nikola Janković



Универзитет у Београду
Филолошки факултет

15.05.2023

Izbor sistema anotacije

- Formalno polazište: Standard za kodiranje korpusa (Corpus Encoding Standard, CES) – saradnja projekata EAGLES, MULTEXT i Vassar/CRNS
- Osnovni kriterijumi Standarda za kodiranje korpusa:
  1. Obradivost (processability)
     - Širok spektar anotiranih jezičkih fenomena na fonetsko-fonološkom, morfološkom, sintaksičkom i diskursnom planu, uključujući i nelingvističke i paralingvističke elemente
     - Morfosintaksička anotacija pomoću alata „TreeTagger“
     - Međusobno preklapanje i upućivanje između elemenata
     - Neophodan opsežan i prilagodljiv sistem anotacije
  2. Mogućnost validacije (validatability)
     - Definisanje formalnih strukturnih specifikacija sa kojima svaki fajl mora biti usaglašen
     - XML format i DTD fajl za validaciju

Izbor sistema anotacije

3. Doslednost
   - Određivanje glavnih strukturnih elemenata teksta
   - Definisanje sistema koji može da predstavi navedene elemente, njihove atribute i veze između njih zadovoljavajući pritom prethodno pomenute zahteve
   - **TEI (Text Encoding Initiative) smernice**
   - Kreirana nova šema prema zahtevima ITALSERB korpusa pomoću TEI-Roma alata

- Drugi razmotreni sistemi anotacije
  - CES
    - Nedovoljno širok i detaljan sistem za potrebe ITALSERB korpusa
  - VALICO-UD
    - Ne zadovoljava kriterijume obradivosti u kontekstu ITALSERB korpusa
  - CHILDES / CHAT format
    - Prilagođen analizama na fonetsko-fonološkom nivou i analizama razvoja jezika dece

Text Encoding Initiative (TEI)

- Konzorcijum koji razvija i održava standard za kodiranje teksta u digitalnoj formi
- Osnovan 1987. godine
- TEI smernice koje precizno određuju metode kodiranja za mašinski čitljive tekstove
- TEI Roma alat za kreiranje sopstvenih šema za validaciju TEI XML fajlova prema specifičnim zahtevima korpusa
- Neke od organizacija koje su podržale TEI:
  - Modern Language Association (MLA)
  - European Union's Expert Advisory Group for Language Engineering Standards (EAGLES)
  - US National Endowment for the Humanities
  - UK's Arts and Humanities Research Board

Metapodaci (uz svaki tekst)

- Identifikacioni broj teksta
  - Primer: A11_B1_01 – sastoji se od slova A (anno), godine ispitivanja, jezičkog nivoa i rednog broja teksta
- Identifikacioni broj studenta
- Datum ispitivanja (mesec i godina)
- Jezički nivo (A2, B1, B2, ili C1)
- Ispitivani jezički nivoi studenta
- Godina studija (1, 2, 3, ili 4)
- Odabrane teme za razgovor na ispitu
- Ispitivač
- Transkiptor(i)
- Naziv audio-fajla
- Trajanje razgovora

Metapodaci vezani za studente

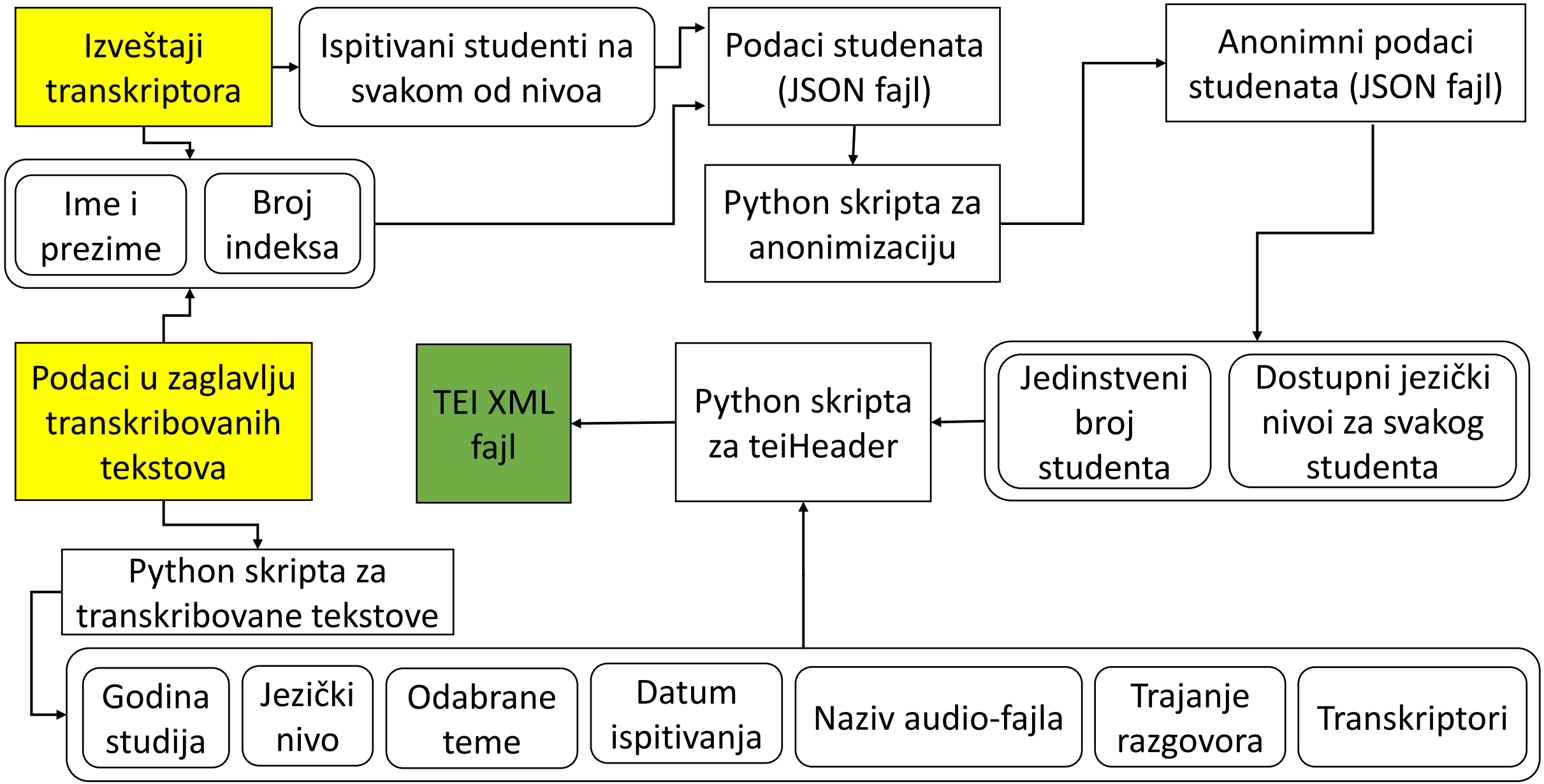
Kompletni podaci učenika (JSON fajl)

- Broj indeksa
- Ime i prezime
- Ispitivani jezički nivoi studenta
- Jedinstveni identifikacioni broj kandidata u korpusu

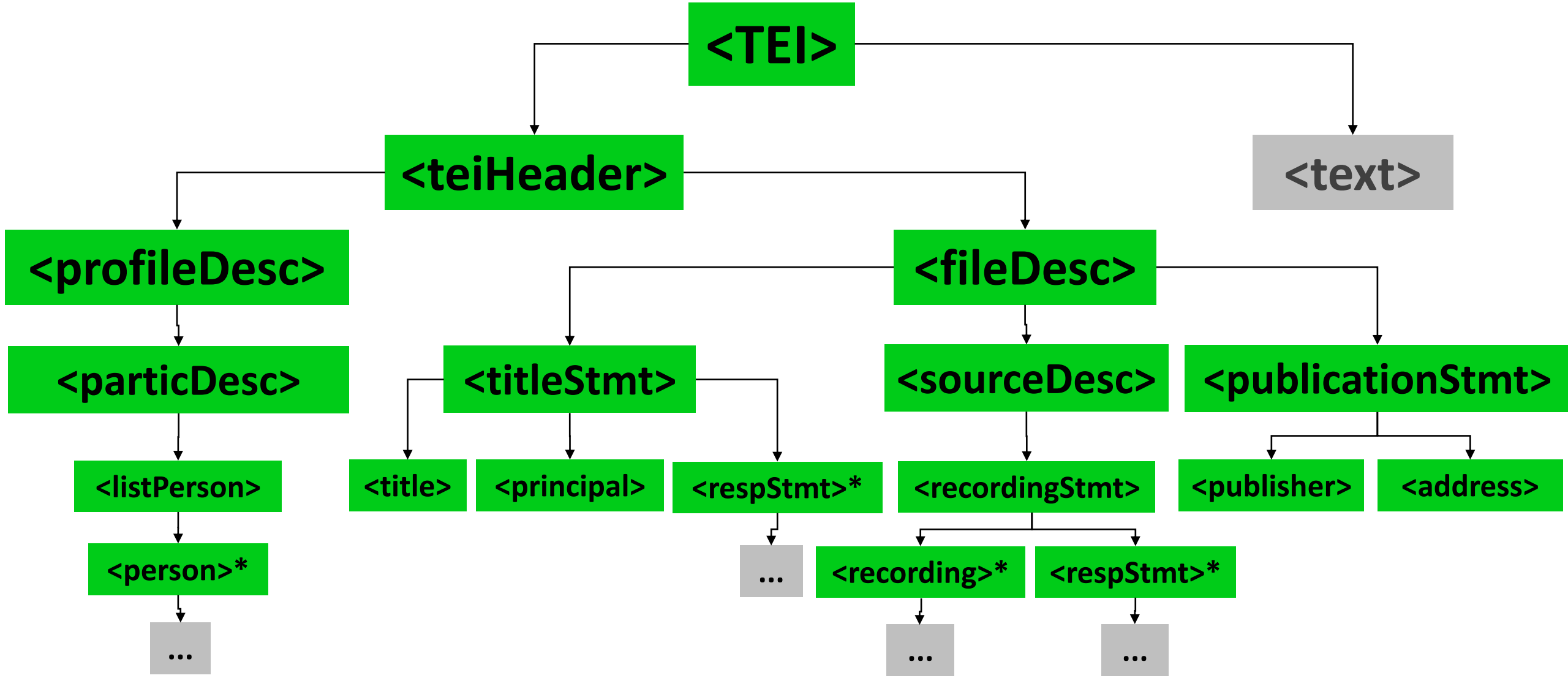
Anonimizirani podaci učenika (JSON fajl)

- Jedinstveni identifikacioni broj kandidata u korpusu
- Ispitivani jezički nivoi studenta

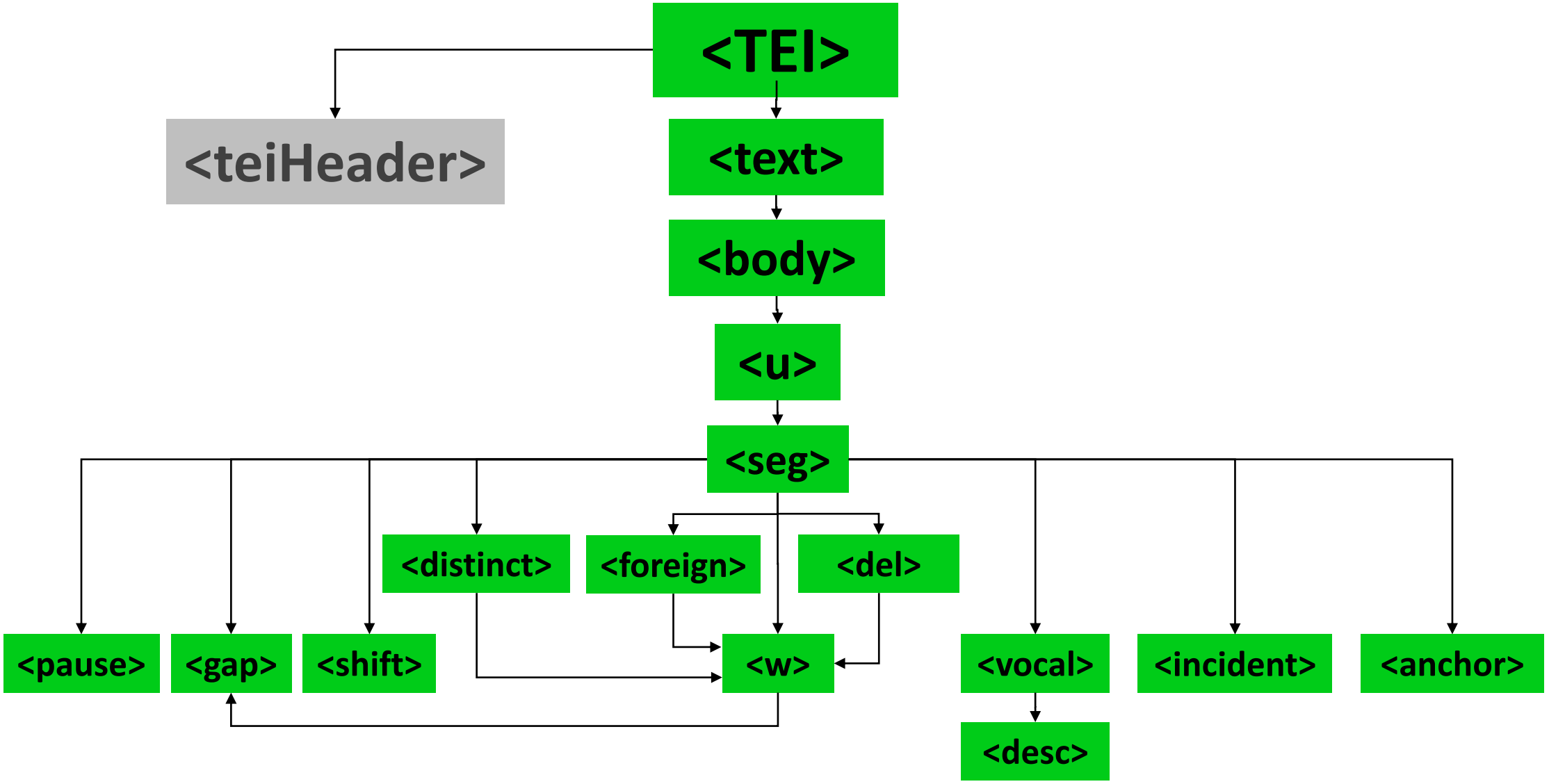
Proces obrade metapodataka



Struktura XML fajlova - zaglavlje



Struktura XML fajlova - tekst



Prilagođavanje prethodne transkripcije i anotacije

- Anotacija u nivou sa ortografskom transkripcijom u originalnom sistemu (zasnovanom na sistemu anotacije CLIPS korpusa)
- CLIPS sistem anotacije ne koristi XML tagove (<>)
- Potreba za konverzijom pojedinačnih tipova anotacije u XML tagove (ili više njih) i u nekim slučajevima za njihovim međusobnim povezivanjem, u skladu sa TEI smernicama
- Prilagođavanje izvršeno pomoću Python skripta uz zastupljenu funkciju regularnih izraza

Promene u slojevima anotacije

- Dodati nivoi anotacije:
  - PoS-tagovi (anotacija prema vrsti reči)
  - Podaci o lemi
  - Jedinstveni identifikator svakog tokena i elementa
  - Formalno definisane veze između povezanih (ili vremenski preklopljenih) elemenata
- Izostavljena anotacija iz prethodnog sistema
  - Ispunjene pauze sa produženim vokalima izjednačene sa drugim ispunjenim pauzama (<e<ee>eh> = <eeh>)
  - Anotacija potvrde prijema (back-channels) sa uzvikom izjednačena sa onima bez uzvika (<ahah!> = <ahah>)

Proces morfosintaksičke anotacije

- Alat za anotaciju „Treetagger“
  - Treetagger parametri za italijanski jezik: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian.par.gz
  - Italijanski tagset:

Tag	Značenje
ABR	abbreviation
ADJ	adjective
ADV	adverb
CON	conjunction
DET:def	definite article
DET:indef	indefinite article
FW	foreign word
INT	interjection

Tag	Značenje
LS	list symbol
NOM	noun
NPR	name
NUM	numeral
PON	punctuation
PRE	preposition
PRE:det	preposition+article
PRO	pronoun

Tag	Značenje
PRO:demo	demonstrative pronoun
PRO:indef	indefinite pronoun
PRO:inter	interrogative pronoun
PRO:pers	personal pronoun
PRO:poss	possessive pronoun
PRO:refl	reflexive pronoun
PRO:rela	relative pronoun
SENT	sentence marker
SYM	symbol
VER:cimp	verb conjunctive imperfect
VER:cond	verb conditional

Tag	Značenje
VER:cpre	verb conjunctive present
VER:futu	verb future tense
VER:geru	verb gerund
VER:impe	verb imperative
VER:impf	verb imperfect
VER:infi	verb infinitive
VER:pper	verb participle perfect
VER:ppre	verb participle present
VER:pres	verb present
VER:refl:infi	verb reflexive infinitive
VER:remo	verb simple past

Proces morfosintaksičke anotacije

1. Razdvajanje ortografske transkripcije od ostalih slojeva anotacije
   - Označavanje turnusa kao osnovne jedinice teksta (<u>)
   - Prepoznavanje i konverzija prethodne anotacije u okviru turnusa i njeno prevođenje u oblik SGML tagova (< >), koji su sada razdvojeni od ortografske transkripcije
   - Dodeljivanje jedinstvenih identifikatora turnusima i anotiranim elementima
   - Označavanje nižih strukturnih jedinica teksta (<seg> i <w>)
   - Povezivanje relevantnih XML elemenata putem atributa „corresp“ i „sync“
2. Međukorak ubacivanja tagova sa informacijama u obliku teksta (#PCDATA) u dodatni tag kako bi bili ignorisani od strane Treetaggera
3. Pokretanje morfosintaksičke anotacije i lematizacije u alatu „Treetagger“
   - Odabrana opcija ignorisanja SGML tagova

Proces morfosintaksičke anotacije

4. Ispravke grešaka u morfosintaksičkoj anotaciji pomoću regularnih izraza
5. Ubacivanje podataka o vrsti reči i lemi u atribute „pos“ i „lemma“ za svaku reč (<w>)
6. Brisanje „zaštitnih“ tagova iz međukoraka
7. Dodeljivanje jedinstvenih identifikatora <seg> i <w> elementima
8. Spajanje zaglavlja (teiHeader) i teksta u TEI XML fajl
9. Validacija na osnovu DTD fajla kreiranog pomoću TEI „Roma“ alata

Prilagođavanje prethodne transkripcije i anotacije

Simbol (INTERPUNKCIJA)	Primena	Primer	XML anotacija
. (tačka)	Sint.-seman. granica	*Sì.*	<seg type="declarative"> <w> sì </w> </seg>
, (zarez)	Sint.-semant. granica	*Sì, a volte quando...*	<w>,</w>
? (upitnik)	Upitni iskaz	*E la sera?*	<seg type="interrogative"> <w>...</w>* </seg>
! (uzvičnik)	Uzvični iskaz	*Buongiorno!*	<seg type="exclamative"> <w> buongiorno </w> </seg>
- (crtica)	Alfabet. citiranje	*Un Ci-Di*	
‘ (apostrof)	Afereza, elizija	*Un po’*	<w>un</w> <w>po</w> <w>’</w>
A…, B…, C… (vel. slovo)	Imena, toponimi, naslovi	*Ho Voglia Di Te*	

Simbol (LEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
22 (brojevi)	Slovna transkripcija	*Ho ventidue anni*	
+ (plus)	Fragment reči	*obbli+ (obbligatoria)*	<del type="truncation"> <w> obbli </w> </del>
/ (kosa crta)	Pogrešan start	*un lavoro… che ha / cerca<aa>*	<w>...</w>* <gap reason="false_start" /> <w>...</w>*

Simbol (LEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
_ (donja crta)	Prekid unutar reči	*sa_lario*	<shift xml:id="A11_B1_01_C09013_24_sh_1" feature= "m_w_pause" new="yes" corresp= "A11_B1_01_C09013_24_w_20" /> <w corresp= "A11_B1_01_C09013_24_sh_1" rend="sa_lario" xml:id="A11_B1_01_C09013_24_w_20"> salario </w>
*(asterisk)	Nepostojeće reči	**rifaro*	<sic> <w> rifaro </w> </sic>

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
<sp>, <lp>	Neispunjena, prazna pauza: kratka <sp> ili duga <lp>	*sì sì sì <lp> <eeh> il mio passatempo preferito. <sp> <eeh> <sp> adesso…*	<seg type=„declarative"> <w>*sì*</w> <w>*sì*</w> <w>*sì*</w> *<pause type="non-lexical" subtype= „long" />* *<pause type="semi-lexical" subtype="eeh" />* <w>*il*</w> <w>*mio*</w> <w>*passatempo*</w> <w>*preferito*</w> </seg> <seg> *<pause type="non-lexical" subtype= "short" />* *<pause type="semi-lexical" subtype="eeh" />* *<pause type="non-lexical" subtype= "short" />* <w>*adesso*</w>

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
<eeh>, <ehm>	„Glasna“ pauza ispunjena vokalizacijom ili nazalizacijom	*mi alzo di solito alle <eeh> dieci o dieci e mezza <ehm>*	*<w>mi</w> <w>alzo</w> <w>di</w> <w>solito</w> <w>alle</w> <pause type="semi-lexical" subtype="eeh" /> <w>dieci</w> <w>o</w> <w>dieci</w> <w>e</w> <w>mezza</w> <pause type="semi-lexical" subtype="ehm" />*

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
<vv> (v=vokal), <kk> (k=konsonant)	Produženi izgovor glasova	*Le traduzioni<ii> <ss>sono<oo>*	*<w>le</w>* *<shift xml:id="A11_B1_01_C09013_40_sh_19" feature="phoneme_lenghthening" new="yes" corresp= "A11_B1_01_C09013_40_w_25" />* *<w corresp= "A11_B1_01_C09013_40_sh_19" rend="traduzioni[ii]" xml:id= "A11_B1_01_C09013_40_w_25"> traduzioni</w>* *<shift xml:id="A11_B1_01_C09013_40_sh_20" feature="phoneme_lenghthening" new="yes" corresp= "A11_B1_01_C09013_40_w_26" />* *<w corresp= "A11_B1_01_C09013_40_sh_20" rend="[ss]sono[oo]" xml:id= "A11_B1_01_C09013_40_w_26">sono</w>*

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
<eh>, <ah>, <mh>, <ahah>, <mhmh>	Potvrda prijema (*back-channels*)	*C: … era una gita* *E: <mhmh>*	*<u who="C..." <seg>* *<w>era</w> <w>una</w> <w>gita</w>* *</seg> </u>* *<u who="E..." <seg>* *<vocal> <desc mhmh </desc> </vocal>* *</seg> </u>*
<ah!>, <eh!>	Uzvici	*<tongue click> <eh!>*	*<vocal> <desc> tongue-click </desc> </vocal>* *<vocal> <desc> eh!</desc> </vocal>*

Simbol (NELEKSIČKI ELEMENTI)	Primena	XML anotacija
<breath>, <clear-throat>, <cough>, <inspiration>, <laugh>, <tongue-click>	Vokalne, neverbalne pojave (kašalj, smeh, …)	*<vocal> <desc > breath </desc> </vocal>* *<vocal> <desc > clear-throat </desc> </vocal>* *<vocal> <desc > cough </desc> </vocal>* *<vocal> <desc > inspiration </desc> </vocal>* *<vocal> <desc > laugh </desc> </vocal>* *<vocal> <desc > tongue-click </desc> </vocal>*

Simbol (NELEKSIČKI ELEMENTI)	Primena	d	XML anotacija
{ }	Preklapanje akustičkih pojava s tekstom	*{<NOISE> bene}*	<seg type="multiple_features" xml:id= "A12_B2_28_C09002_18_m_f_1"> <w corresp= "A12_B2_28_C09002_18_m_f_1"> bene </w> <incident corresp= "A12_B2_28_C09002_18_m_f_1"> <desc corresp = "A12_B2_28_C09002_18_m_f_1"> noise </desc> </incident> </seg>
<i.talkers>	Glasovi drugih govornika u pozadini	{<i.talkers> piacevole}	<seg type="multiple_features" xml:id= "A12_B2_28_C09002_35_m_f_2"> <w corresp= "A12_B2_28_C09002_35_m_f_2" piacevole </w> <incident corresp="A12_B2_28_C09002_35_m_f_2"> <desc corresp = "A12_B2_28_C09002_35_m_f_2> i_talkers </desc> </incident> </seg>

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
<NOISE >	Prateće akustičke pojave	*<NOISE> che tipo sei?*	<seg type="multiple_features" xml:id= "A12_B2_28_C09002_18_m_f_1"> <w corresp= "A12_B2_28_C09002_18_m_f_1" che </w> <w corresp= "A12_B2_28_C09002_18_m_f_1" tipo </w> <w corresp= "A12_B2_28_C09002_18_m_f_1" sei </w> <incident corresp="A12_B2_28_C09002_18_m_f_1"> <desc corresp = "A12_B2_28_C09002_18m_f_1"> noise </desc> </incident> </seg>
<unclear>	Nerazumljiva reč ili sekvenca	*mi piace molto <unclear>*	*<w><gap reason="unclear" /> </w>*

Simbol (NELEKSIČKI ELEMENTI)	Primena	Primer	XML anotacija
#	Preklapanje turnusa	E: di conferma #<C> diciamo# C: #<E> sì, di conferma#	<u who="E02" xml:id="A12_B2_26_E02_33"> <seg> <w> di </w> <w>conferma</w> </seg> <seg type="overlap" xml:id="A12_B2_26_E02_33_seg_55"> <anchor xml:id="A12_B2_26_E02_33_t_9a" synch="A12_B2_26_C09036_34_t_9b" /> <w>diciamo</w> <anchor xml:id="A12_B2_26_E02_33_t_10a" synch="A12_B2_26_C09036_34_t_10b" /> </seg> </u> <u who="C09036" xml:id="A12_B2_26_C09036_34"> <seg type="overlap" xml:id= "A12_B2_26_C09036_34_seg_57"> <anchor xml:id="A12_B2_26_C09036_34_t_9b" /> <w> sì </w> <w>,</w> <w>di</w> <w>conferma</w> <anchor xml:id="A12_B2_26_C09036_34_t_10b" /> </seg> </u>

Simbol (KOMENTARI TRANSKRIPTORA)	Primena	Primer	XML anotacija
[whispering]	Varijacije u tonu glasa	{[whispering] non so}	<seg type="multiple_features" xml:id="A12_B2_29_C09006_28_seg_48"> <shift feature="vocal" new="whisp" corresp="A12_B2_29_C09006_28_seg_48"/> <w corresp = "A12_B2_29_C09006_28_seg_48">non </w> <w corresp = "A12_B2_29_C09006_28_seg_48">so</w> </seg>

Simbol (KOMENTARI TRANSKRIPTORA)	Primena	Primer	XML anotacija
[dialect]	Reč iz dijalekta	*tanto [dialect]*	*<distinct type= "dialect">* *<w> tanto </w>* *</distinct>*
[foreign word]	Strana reč	*ukus [foreign word]*	*<foreign>* *<w>Ada</w>* *</foreign>*

Izgled zaglavlja u programu „TXM“

A11_B1_01_TXT

id	A11_B1_01_txt
cassetta	R_VN780141
durata	0:2:28
data	Maggio 2011
anno-di-studio	2
livello	B1
sede	Belgrado
codice-esami	1
candidato	C09013
livelli-del-candidato	A2:2010 B1:2011 B2:2012 C1:2013
esaminatore	E02

Izgled teksta u programu „TXM“

A11_B1_01_TXT

id A11_B1_01_txt

va bene, XXXXXX, dunque, prova numero uno, quale argomento hai scelto

se hai una piccola somma di denaro da spendere, verso che cosa preferisci orientarti

mhmh okay

inspiration siccome mi piace molto viaggiare spenderei i soldi per esempio ad un viaggio in Italia Mi piacerebbe molto andare a vedere la Sicilia

mhmh

perché ho sentito le cose molto belle Ho Sono già stata a Roma, mi piace molto e anche vorrei tornare a Roma

mhmh mhmh

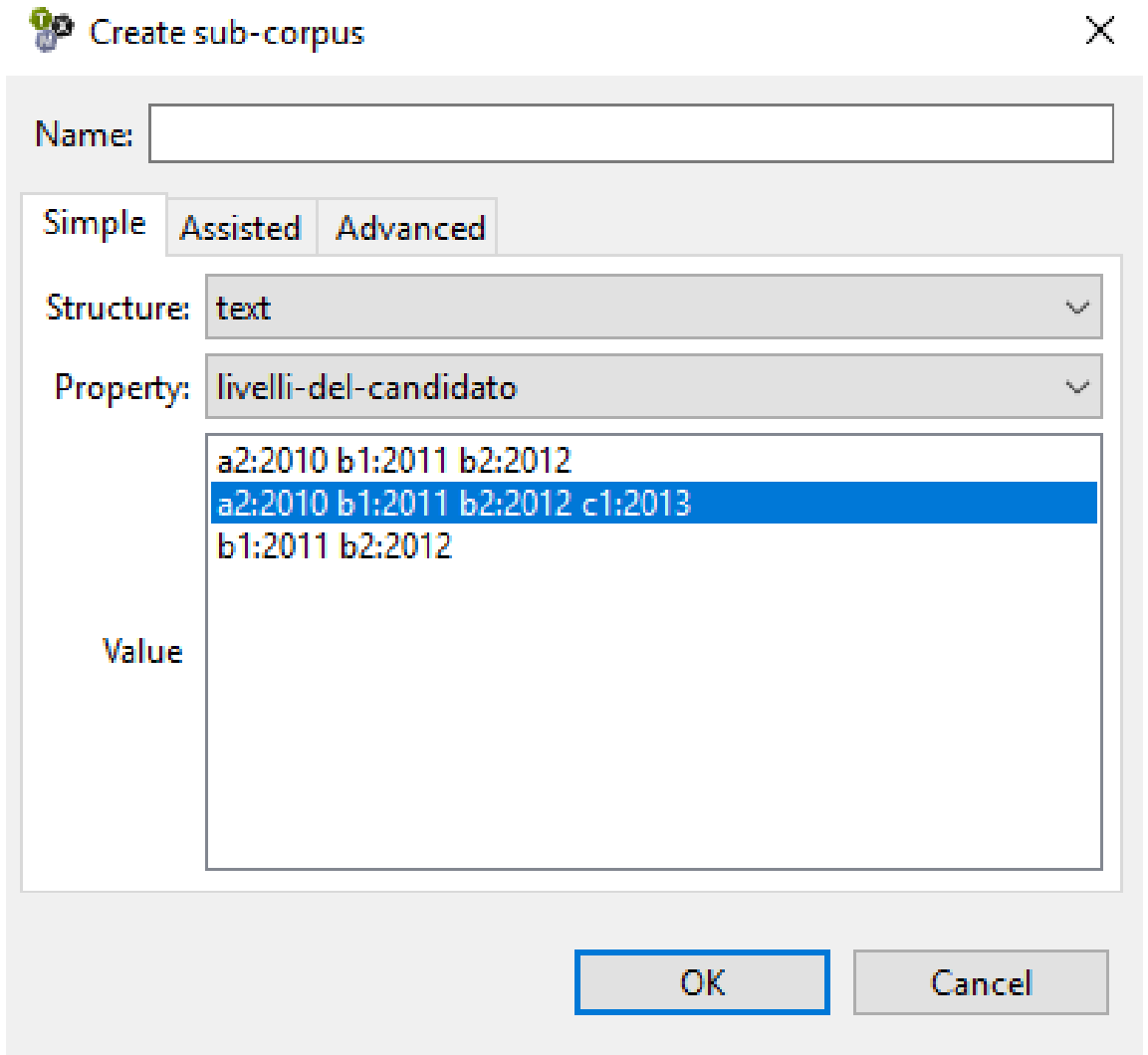
comprerei alcuni libri comprerei alcun pezzo di vestito di abbigliamento

mhmh

non so, andrei al cinema o

Primer upita u korpusu pomoću alata“TXM“

- Pravljenje potkorpusa na osnovu metapodataka – studenti koji su ispitivani na sve četiri godine studija

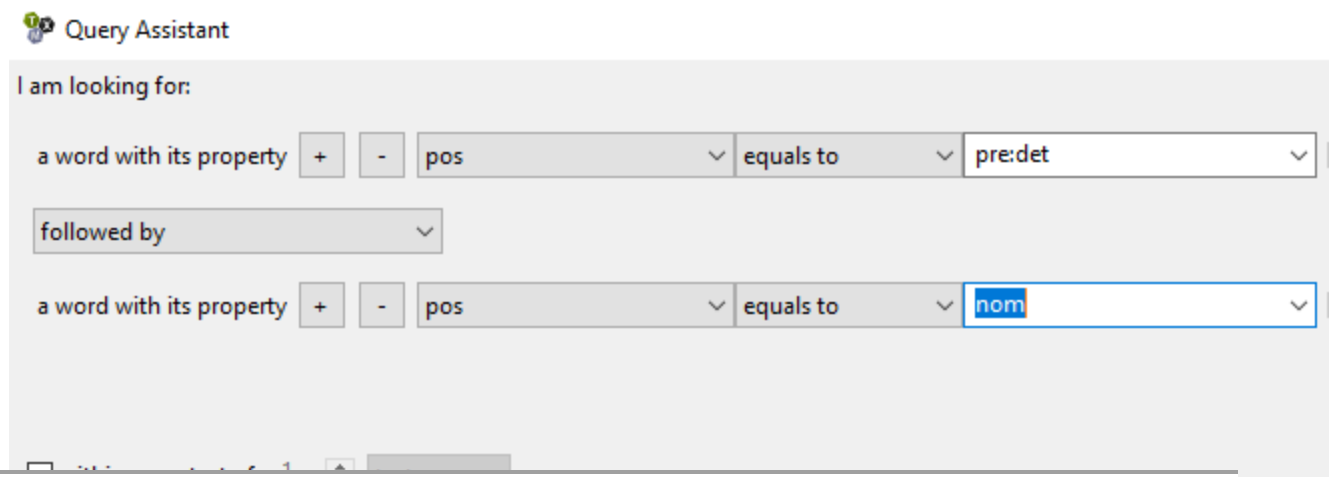


- Konkordance imenice „città“

Query [word = "città"]

text_id	Left context	Pivot	Right context
A11_B1_01	famosi mhmh abbiamo passeggi...	città	italiane sì, sono stata a Venezia, a ...
A11_B1_03	loro vesti mhmh vestito mhmh è ...	città	oppure sì mhmh penso che sia B...
A11_B1_07	e visitare gli visitare i grandi le gra...	città	di di Europa mhmh non so vorrei ...
A11_B1_07	che in Belgrado bene, e quale qu...	città	per esempio le piacerebbe visitar...
A11_B1_07	del sud mhmh e Roma perché è ...	città	vecchia e s __UNDEF__ senza è il p...
A11_B1_07	Italia mhmh in Italia ha visitato gi...	città	inspiration no, no, sono sono stat...
A11_B1_07	mhmh tre volte ma non ho visitat...	città	grande mhmh perché mio papà l...
A11_B1_07	Milan a Milano mhmh perché pe...	città	è più bella del nord mhmh e che ...
A11_B1_09	si torna in tongue-click nel suo n...	città	natale dopo qualche periodo mh...
A11_B1_10	li mhmh e vedere molte cose mh...	città	quali città vorebbe ved __UNDEF_...
A11_B1_10	e vedere molte cose mhmh molt...	città	vorebbe ved __UNDEF__ vedere p...
A11_B1_12	una parte all ' altra nell ' una	città	mhmh non possiamo parlare dell...
A11_B1_13	mi è piaciuto Milano è un __UND...	città	di moda e molto bella e ma non ...
A11_B1_15	due sorelle e a Kralievo è una picc...	città	e anche in Kralievo s __UNDEF__ p...

- Predlozi sa članom ispred imenice:



Query

```
@[pos = "pre:det"] [pos = "nom"]
```

text_id	Left context	Pivot	Right context
A11_B1_01	vestito di abbigliamento mhmh n...	[al] cinema	o mhmh a un ristorante con gli a...
A11_B1_01	Padova okay, okay, bene, passiamo	[alla] prova	numero due laugh laugh hm qua...
A11_B1_01	sole o nuotare mhmh mi piace a...	[al] cinema	o a teatro mhmh mi piace ascolta...
A11_B1_02	va bene, XXXXXX, dunque passia...	[alla] prova	numero uno, quale argomento h...
A11_B1_02	io parlo di un sogno che vorrei re...	[al] primo	posto vorrei finire la facoltà mhm...
A11_B1_02	va bene moderni e poi ve ne parli...	[alla] prova	numero due laugh __UNDEF__ sì l...
A11_B1_02	bene, okay, grazie, dunque passia...	[alla] prova	numero due sì quale argomento ...
A11_B1_02	ha il suo ragazzo ah! ma laugh	[alla] fine	loro due sono stati insieme va be...
A11_B1_03	__UNDEF__ divento nervosa mhm...	[dal] centro	abito a Nuova Belgrado mhmh q...
A11_B1_03	autob ahah dunque, però si arriva...	[al] centro	non è un problema sì, sì no mhm...

- Prikazivanje fragmentiranih reči

Query [_.del_type = "truncation"]

text_id	Left context	Pivot	Right context
A11_B1_01	sola in Sicilia o comunque a viag...	co	__UNDEF__ no con qualche amica...
A11_B1_02	fare l ' interprete i libri mhmh i	roman	__UNDEF__ i romanzi i romanzi sì ...
A11_B1_02	ecco sì ho letto qualche libro itali...	l	__UNDEF__ la prima era di Fabio V...
A11_B1_03	altra scelta che che autobus mh...	m	__UNDEF__ pia non mi piace anda...
A11_B1_03	folla e è molto caldo in autobus e	t	__UNDEF__ divento nervosa mhm...
A11_B1_03	, no no no no, non ha	mai	__UNDEF__ s __UNDEF__ s __UNDE...
A11_B1_03	no no no, non ha mai __UNDEF__	s	__UNDEF__ s __UNDEF__ due mesi...
A11_B1_03	no, non ha mai __UNDEF__ s __U...	s	__UNDEF__ due mesi fa quan __U...
A11_B1_03	__UNDEF__ s __UNDEF__ s __UNDE...	quan	__UNDEF__ quando visitavo mia a...
A11_B1_03	gente sta aspettando autobus o ...	s	__UNDEF__ stanno siedendo alcu...
A11_B1_03	mhmh tram alcuni s __UNDEF__ s...	s	__UNDEF__ stanno in piedi mhmh...
A11_B1_04	, non sono stata purtroppo no m...	i	__UNDEF__ no, Inghilterra mhmh ...
A11_B1_04	cultura e tutti gli aspetti della cult...	so	__UNDEF__ il cibo mhmh voglio p...
A11_B1_04	alla fine lui scopre che suo padre ...	sequre	__UNDEF__ sequestrato un ragazzi...
A11_B1_05	America mhmh dopo un tempo ...	s	__UNDEF__ è andato per otto gior...

Naredni koraci

- Anotacija preostalih transkribovanih tekstova
- Ručno ispravljanje anotacije vezane za vrste reči i leme
- Stilizovanje prikaza elemenata teksta (XSLT/CSS)
- Postavljanje korpusa na Internet i omogućavanje pristupa istraživačima

Reference

- Ceković-Rakonjac, N. (2013). ITALBEG corpus parlato di italiano L2. *Italica Belgradensia*, 1, 336-348.
- Ceković-Rakonjac, N. (2012). Ortografska transkripcija govornog korpusa ESNAKIT. U A. Vraneš, LJ. Marković & G. Alexander (prir.), *Digitalizacija kulturne i naučne baštine, univerzitetski repozitorijumi i učenje na daljinu*, *knj*. *3* (str. 163-182). Beograd: Filološki fakultet. [https://www.academia.edu/8834565 6]
- Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference (LREC),* Granada, Spain, 463-470
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- TEI Consortium, eds. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 4.6.0. Last updated on 4th April 2023. TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Hvala na pažnji!

nikolajankovickv@gmail.com