Syntactic Annotation of Slovene CMC: First Steps

Špela Arhar Holdt*♦, Darja Fišer*‡, Tomaž Erjavec‡, Simon Krek‡

* Faculty of Arts, University of Ljubljana

♦Institute for Applied Slovene Studies Trojina

‡ Jožef Stefan Institute

spela.arharholdt@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si, simon.krek@ijs.si

Introduction

- The annotation of 200 Slovene tweets with the JOS dependency model (Erjavec et al., 2010).
- "Resources, Tools and Methods for the Research of Non-Standard Internet Slovene" (J6-6842, 2014–2017, leader dr. Darja Fišer).
- Results will serve as the groundwork for syntactic annotation and linguistic analysis of Slovene CMC.
- (Linguistic) motivation: no corpus-based description of the syntax of Slovene CMC.

Kons1 from the Janes training corpus (4.000 tweets).

ORIGINAL TWEET

@merineseri pa če ni snicekrsaaaa ;) zdej sm se spomnu, da mamo doma doma arašide, k jih je treba res kopat iz zemlje ;)

Lemmatisation and POS-tagging (Erjavec et al., 2005; Ljubešić et al., 2014).

Manual correction of the sentence segmentation and tokenization. Normalisation on the lexical and morphological level (Čibej et al., 2016a).

NORMALISED TWEET

@merineseri pa če ni Snickersa ;) zdaj sem se spomnil, da imamo doma doma arašide, ki jih je treba res kopati iz zemlje ;)

Manual correction of the attributed lemmas and POS-tags (Čibej et al., 2016b).

Dataset for Syntactic Annotation

- 200 tweets (475 sentences).
- Only tweets longer than 120 characters by private individuals.
- Sampled to include an equal amount of linguistically and technically standard and non-standard tweets (Ljubešić et al., 2015).

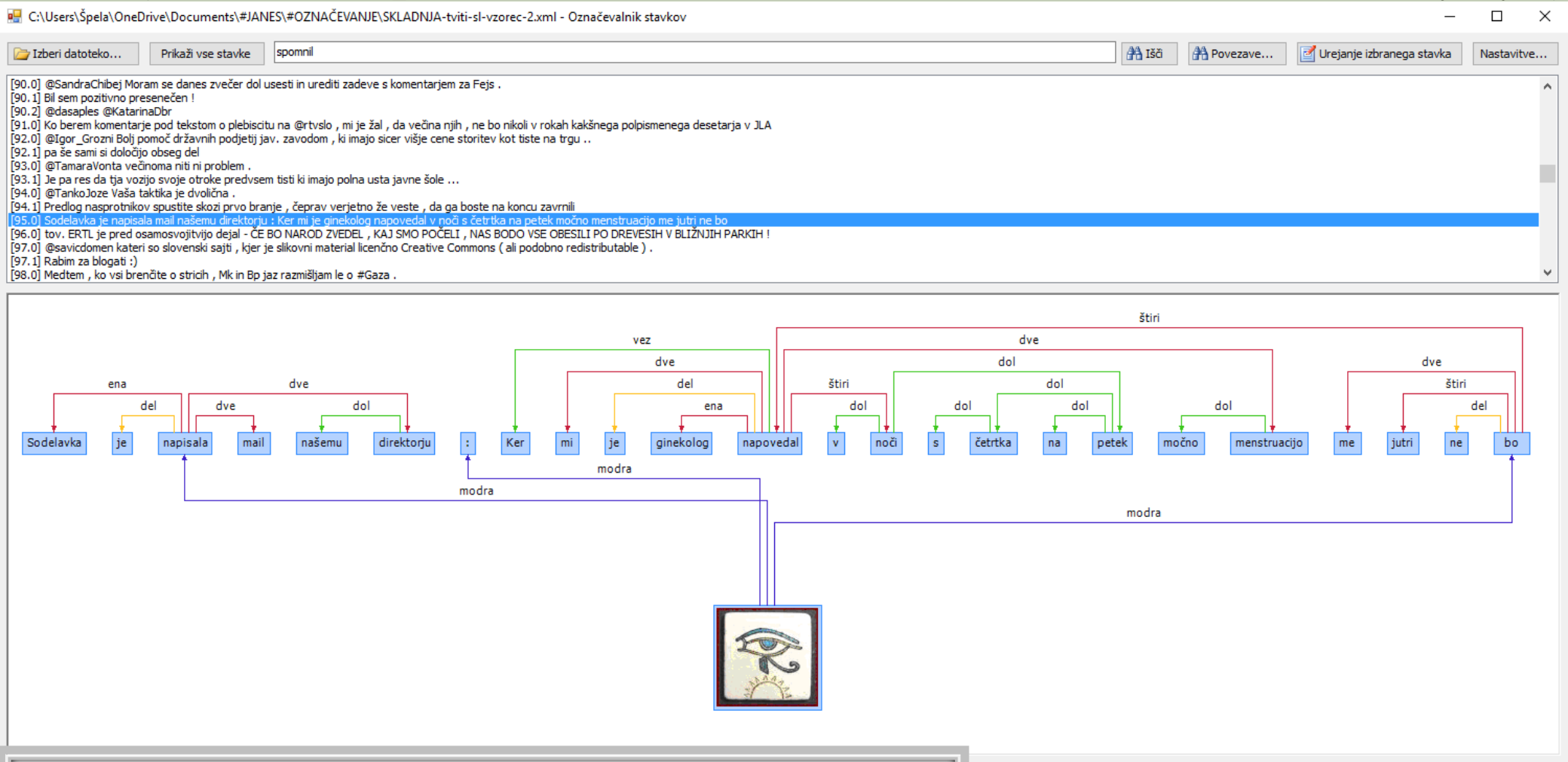
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		id	name	sex	text	created	favorited	retweeted	std_tech	std_tech_n	std_ling	std_ling_n	sentiment	source
2	*	tid.7487	agencija	male	Manj kot dan je trajalo, da smo dobili prvi jailbreak za Applovo najnovejšo različico operacijskega sistema iOS, 4.2.1. Kdo bi si mislil.	2010-11-24T17:36:11	0	0	T1	42095	L1	42156	negative	private
3	*	tid.2661	_MarkoG_	male	Nekateri zvesti podporniki... Še vam ni jasno, da če bi želeli videti vsak tweet kandidatov, bi enostavno sledili njim? #predsednik12	2012-11-07T10:18:51	0	0	T1	42095	L1	42064	negative	private
4	*	tid.3118	markokastelic	male	@petrasovdat V priemerjavi s svojim predhodnikom nedvomno. Okoliščine in pogoji dela pa so mu bili vse prej kot naklonjeni.	2013-03-13T13:58:29	0	0	T1	42095	L1	42125	negative	private
5	*	tid.3360	evabelka	female	Delam, delam, delam, odstranil bom plevel, prekopal bom vrtiček, prepeval ves vesel... #gardening inspired by Palček Primož #nowsinging	2013-05-19T09:44:09	1	0	T1	42095	L1	42095	positive	private
6	*	tid.3427	PeterSuhel	male	Čeferin: sodišče podlega javnemu mnenju. Ni samo obsodilna sodba tista, ki kaže na to, da pravna država funkcionira. #pogledislovenije	2013-06-06T19:02:39	1	3	T1	42005	L1	1.0	negative	private
7	*	tid.3526	Pizama	male	Lažji med nogami: Osme #Glave s pregledom 3. sezone #GoT. Gobcamo @anzet @BokiNachbar @WiC_HmR @matevzluzar	2013-07-04T07:32:31	2	4	T1	1.0	L1	1.0	neutral	private
8	*	tid.3614	AleksHribovsek	male	Pozor, ob cesti pri Blagovici je ustavljen lažno okvarjen romunski kombi; ustavljajo nič hudega sluteče naivne voznike na pomoč.	2013-07-28T14:47:44	1	19	T1	1.0	L1	42005	negative	private
9	*	tid.3661	IrenaSirena	female	@peter_pec Plus, premikanje na SD kartico sem probala že stokrat. Tudi telefon to ponudi kot možnost. Rezultat? "Ni predmetov za premikanje"	2013-08-10T11:17:47	0	0	T1	42095	L1	42036	positive	private
10	*	tid.3748	SanjaLT	female	@kricac; Bom ugibala - pripadnost? Današnja mladina tako zelo hlepi po tem. Mi pa tudi verjetno nismo bili tako zelo drugačni. @markopotrc	2013-09-03T13:15:24	0	0	T1	42095	L1	42005	neutral	private
11	*	tid.3817	TyraTweet	female	@LakovicJaka hvala za vse kar si naredil za reprezentanco.Čeprav ti letos ni šlo brez tebe ne bi bili #junaki. Rečem ti lahko le SREČNO.	2013-09-22T10:53:04	1	0	T1	42156	L1	42156	positive	private

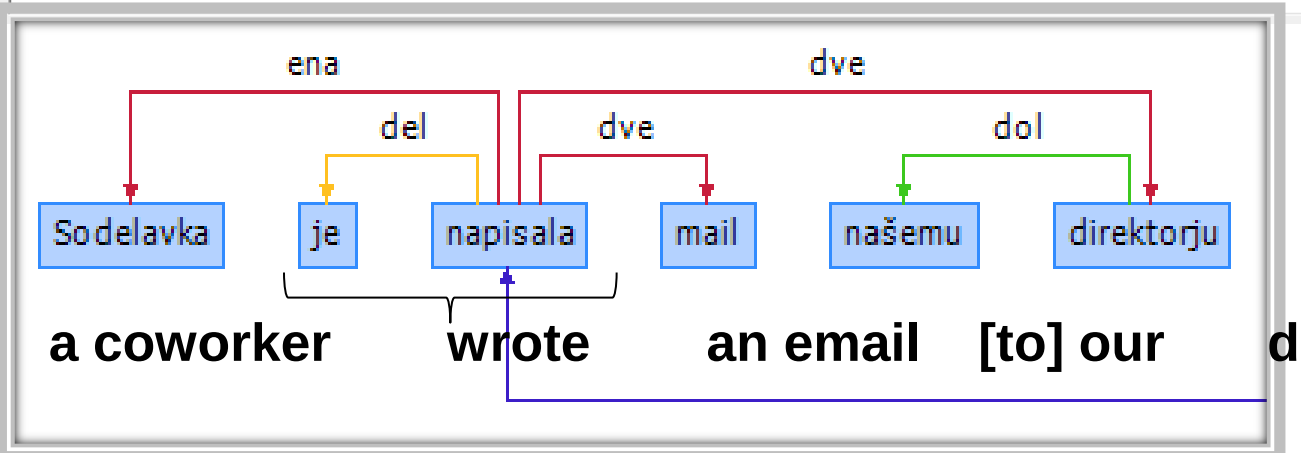
The JOS Dependency Model

- Designed in the project “Linguistic Annotation of Slovene” (Erjavec et al., 2010).
- Applied in the “Communication in Slovene” project to annotate the ssj500k training corpus (Krek et al., 2015), on the basis of which a parser for Slovene was trained (Dobrovoljc et al., 2012).
- Based on syntactic dependencies; simpler and more robust than similar models.
- A specialised program for the visualisation, manual annotation and search of the data was developed (J. Brank).

The JOS Dependency Model

Groups of labels	Labels	Description
First level labels link elements in different types of phrases (green and yellow colour in the visualisation).	dol	Links heads and modifiers in phrases.
	del	Links parts of verbal phrases.
	prir	Links heads in coordinate structures within clauses.
	vez	Links words or commas in conjuctive function.
	skup	Links (function) words in frozen multi-word structures.
Second level labels link sentence elements (red colour in the visualisation).	ena	Clause subject.
	dve	Clause object.
	tri	Adverbial of manner.
	štiri	Other adverbials.
Third level label links all other structures (blue colour in the visualisation).	modra	Links to the root, punctuation, fragments, etc.


C:\Users\Špela\OneDrive\Documents\#JANES\#OZNAČEVANJE\SKLADNJA-tviti-sl-vzorec-2.xml - Označevalnik stavkov
Izberi datoteko...
Prikaži vse stavke
spomnil
Išči
Povezave...
Urejanje izbranega stavka
Nastavitve...
[90.0] @SandraChibej Moram se danes zvečer dol usesti in urediti zadeve s komentarjem za Fejs .
[90.1] Bil sem pozitivno presenečen !
[90.2] @dasaples @KatarinaDbr
[91.0] Ko berem komentarje pod tekstom o plebiscitu na @rtvslo , mi je žal , da večina njih , ne bo nikoli v rokah kakšnega polpismenega desetarja v JLA
[92.0] @Igor_Grozni Bolj pomoč državnih podjetij jav. zavodom , ki imajo sicer višje cene storitev kot tiste na trgu ..
[92.1] pa še sami si določijo obseg del
[93.0] @TamaraVonta večinoma niti ni problem .
[93.1] Je pa res da tja vozijo svoje otroke predvsem tisti ki imajo polna usta javne šole ...
[94.0] @TankoJoze Vaša taktika je dvolična .
[94.1] Predlog nasprotnikov spustite skozi prvo branje , čeprav verjetno že veste , da ga boste na koncu zavrnili
[95.0] Sodelavka je napisala mail našemu direktorju : Ker mi je ginekolog napovedal v noči s četrtka na petek močno menstruacijo me jutri ne bo
[96.0] tov. ERTL je pred osamosvojitvijo dejal - ČE BO NAROD ZVEDEL , KAJ SMO POČELI , NAS BODO VSE OBESILI PO DREVESIH V BLIŽNJIH PARKIH !
[97.0] @savicdomen kateri so slovenski sajti , kjer je slikovni material licenčno Creative Commons (ali podobno redistributable) .
[97.1] Rabim za blogati :)
[98.0] Medtem , ko vsi brenčite o stricih , Mk in Bp jaz razmišljam le o #Gaza .
štiri
vez
dve
dve
dol
dve
ena
dve
del
del
štiri
štiri
dol
dol
dol
dol
dol
del
ena
Sodelavka
je
napisala
mail
našemu
direktorju
:
Ker
mi
je
ginekolog
napovedal
v
noči
s
četrtka
na
petek
močno
menstruacijo
me
jutri
ne
bo
modra
modra
modra


ena
dve
del
dve
dol
Sodelavka
je
napisala
mail
našemu
direktorju
a coworker
wrote
an email
[to] our
director

200 NORMALISED TWEETS WITH CORRECTED TAGS

@merineseri pa če ni Snickersa ;) zdaj sem se spomnil, da imamo doma doma arašide, ki jih je treba res kopati iz zemlje ;)

Automatically parsed (Dobrovoljc et al., 2012) and imported into the SSJ annotation program.

Manual correction of the syntactic annotations.

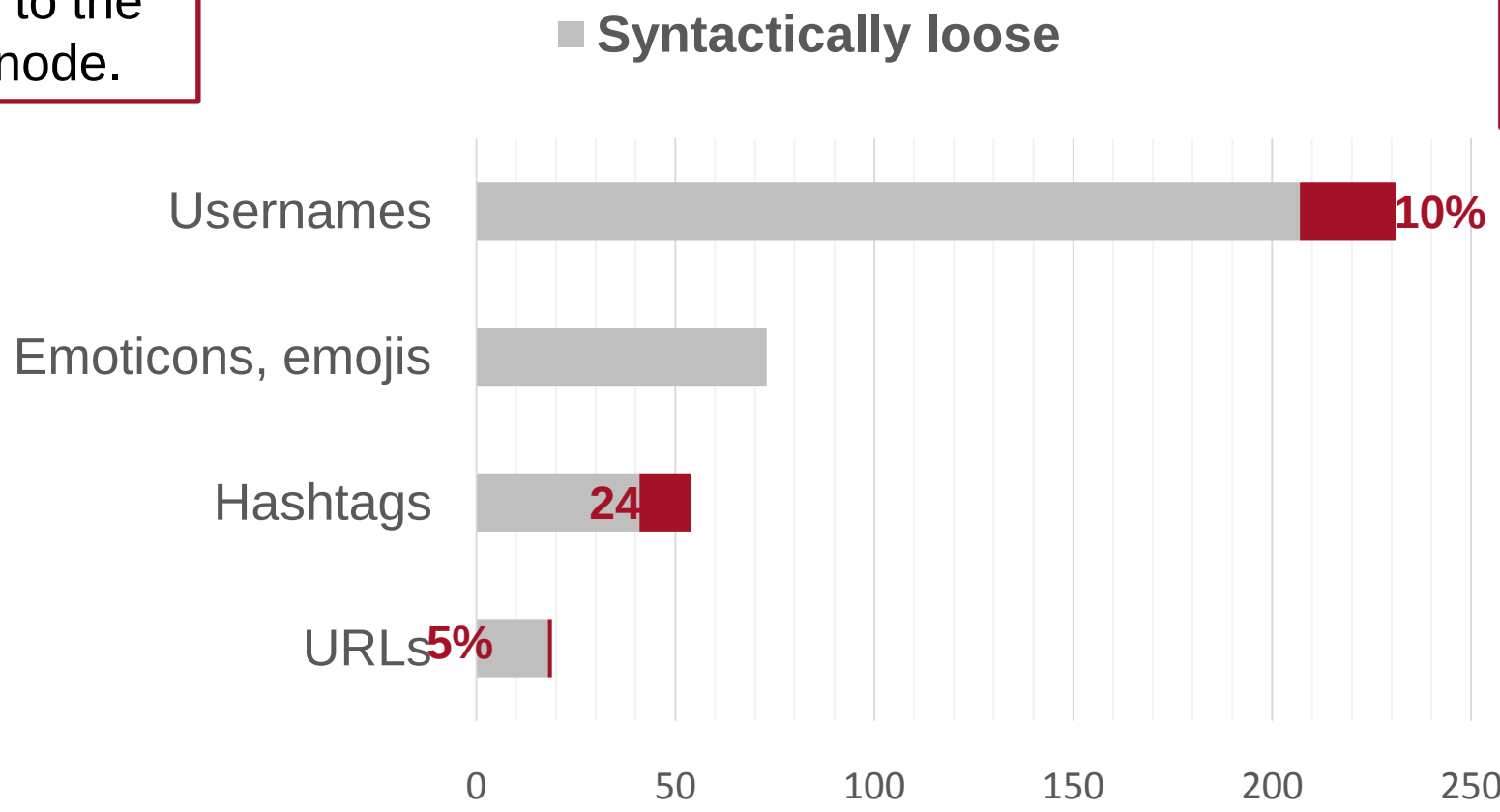
Guidelines that were used for the annotation of standard Slovene (Holozan et al., 2008).

Twitter elements;
foreign language;
non-standard use of punctuation;
syntactical fragments and ellipsis;
other non-standard features.

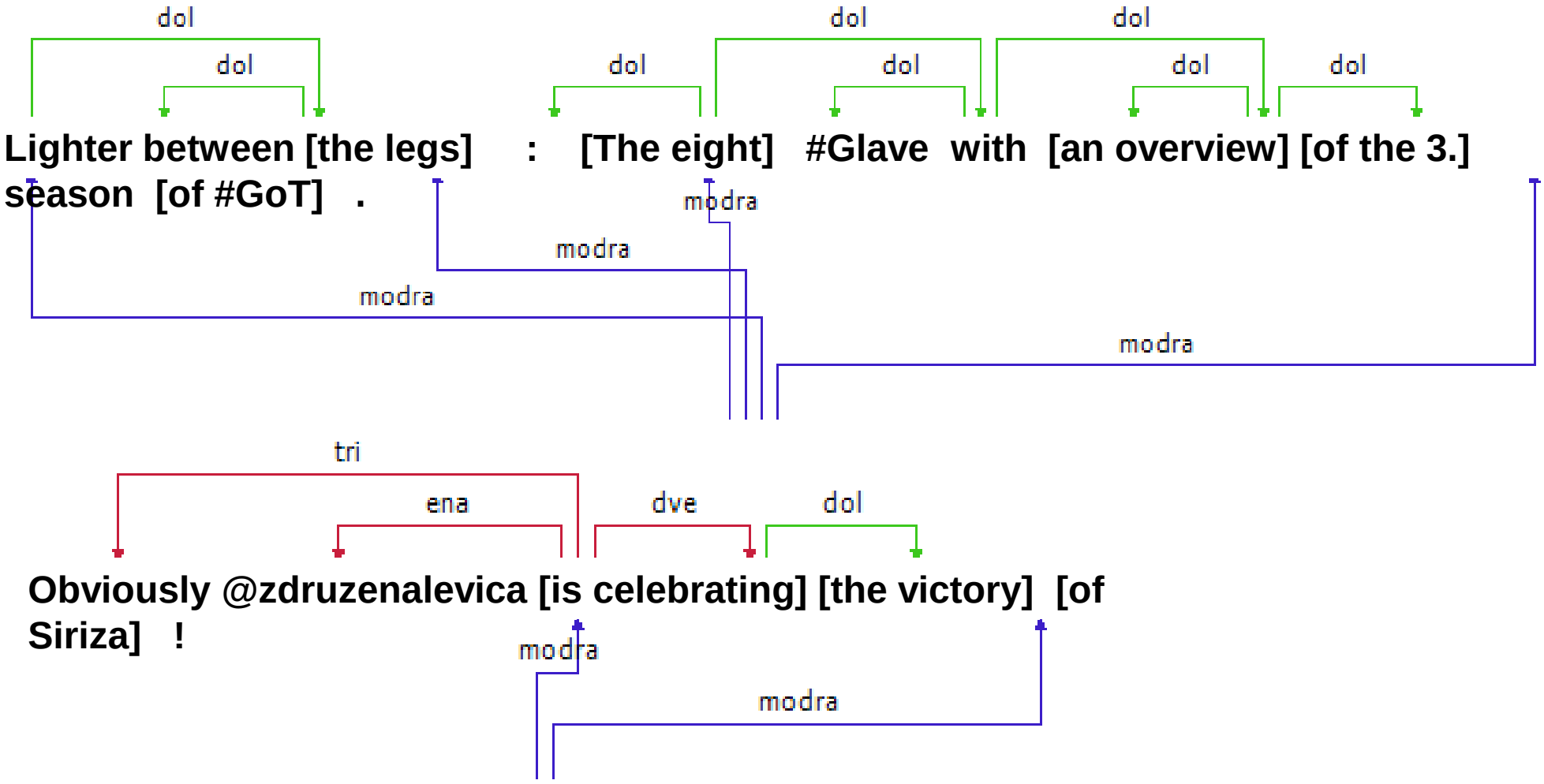
Twitter Elements

Connected to the node.

Annotated according to their function.



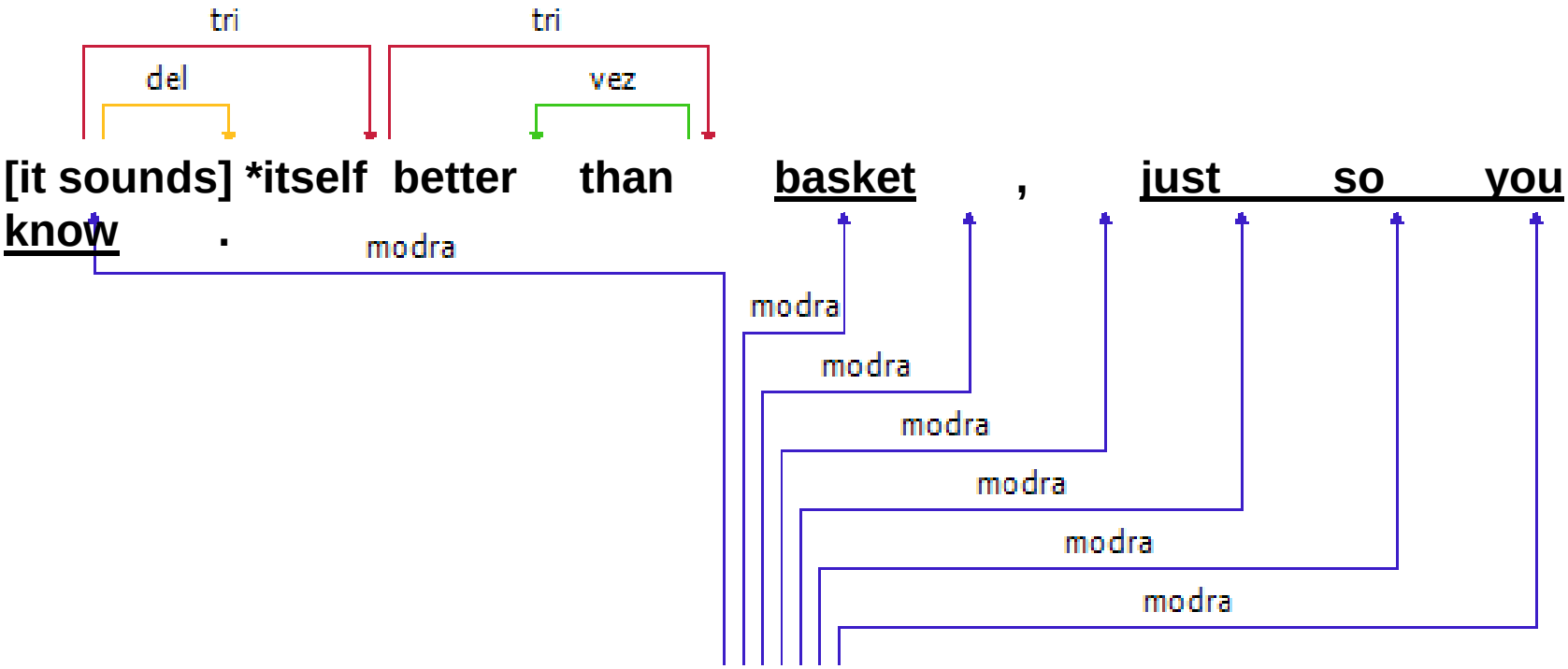
Twitter Elements



Foreign Language

- Appears in 26% of the annotated tweets.
- Primarily from English (20%) and related South Slavic languages (6%).
- Different levels of adaptation to Slovene regarding the spelling and morphology of these elements (Čibej et al., 2016b).
- Single words (41), word phrases (12), or longer segments/clauses (17).
- Single words and two-part phrases with a clear dependency relation are attached into the syntactic tree, whereas in longer phrases and segments, all of the foreign elements get attached to the node.

Foreign Language

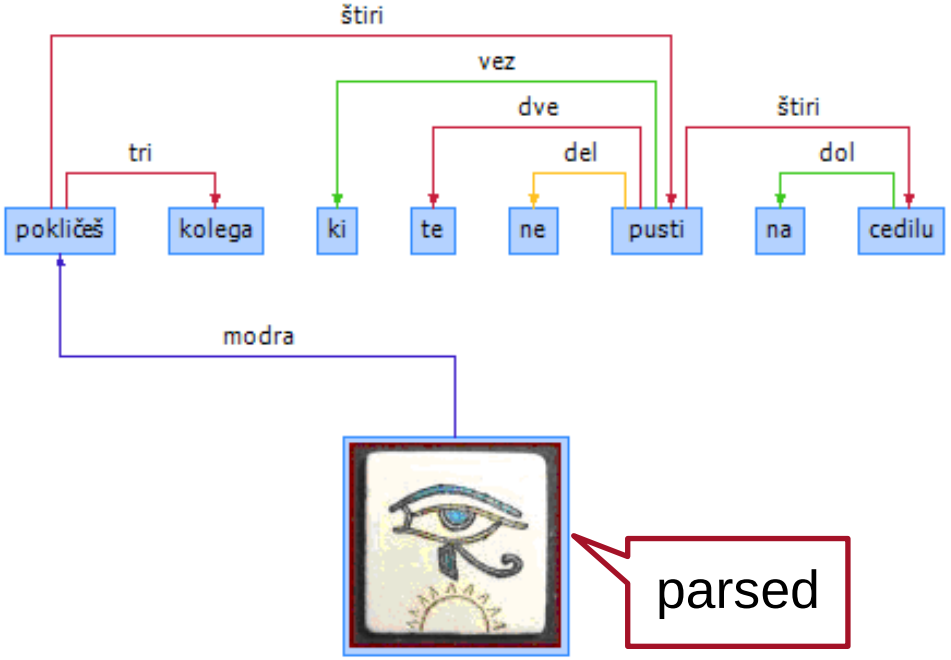


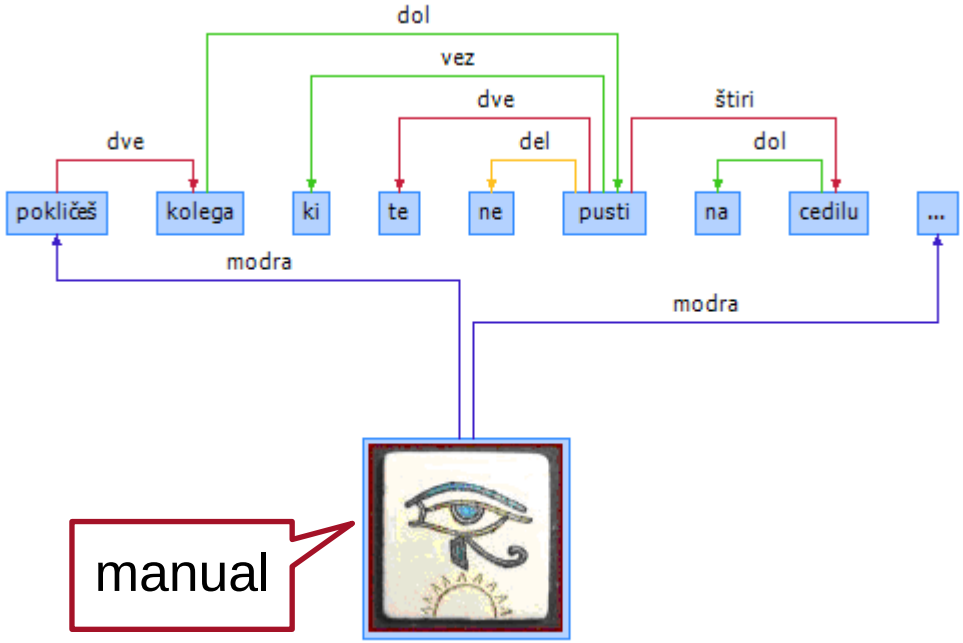
Non-standard Use of Punctuation

- The existing parser for Slovene is trained on standard language, where punctuation – especially the use of the comma – plays an important role in determining the borders between clauses and other types of sentence segments.
- Omitted, redundant, and misplaced commas lead to parsing mistakes. 33% of the annotated tweets exhibit one or more such problems.
- Emoticons as terminal punctuation marks; two or three dots as terminal/internal punctuation

  težave z elektriko … pokličeš kolega ki te ne pusti na cedilu … spiješ enega ali dva … nice saturday :) #winwinsituation #JackDaniels
- Include a step of punctuation normalisation?

Non-standard Use of Punctuation



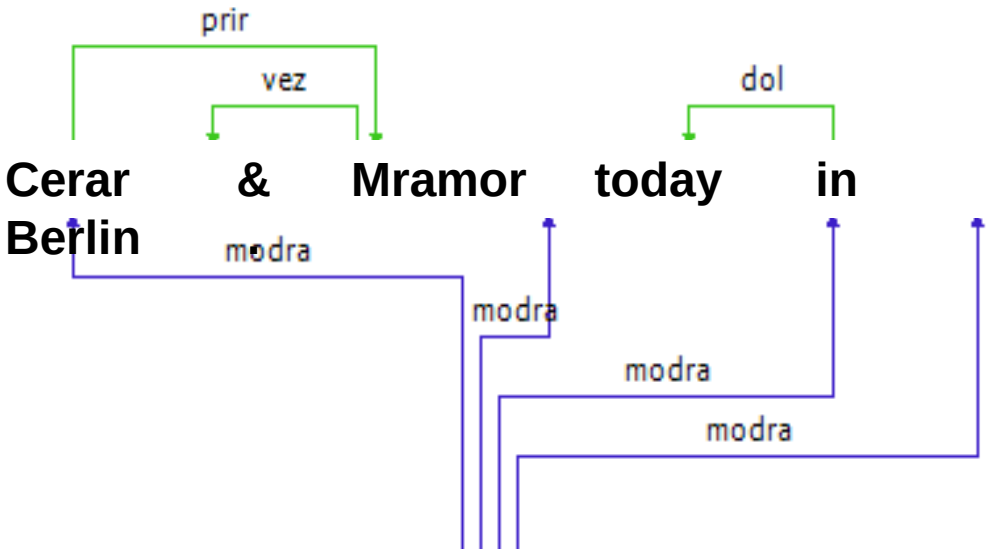


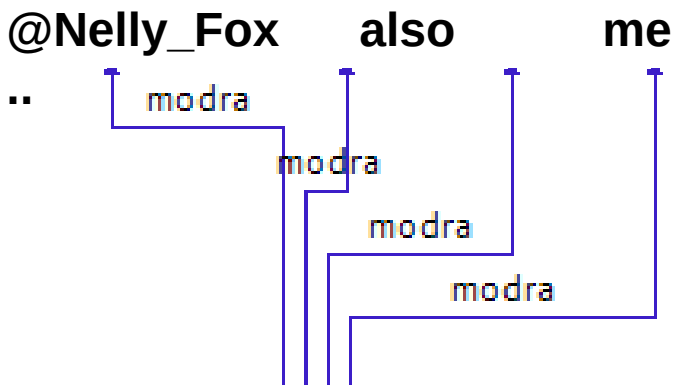
Pokličeš kolega , ki te ne pusti na cedilu.
You call a friend , who doesn‘t let you down.

Syntactical Fragments and Ellipses

- While in regular clauses only the head of the predicate is attached to the root (or the relevant ordinate clause), in fragments or clauses without the predicate, each separate phrase head attaches to the node as well.
- Attaching to the root speeds up the annotation process (Kong et al., 2012).
- Alternative systems that allow for the orphan node to be promoted to the place of the missing parent (Dobrovoljc and Nivre, 2016).

Syntactical Fragments and Ellipses





Other Non-standard Language Features

- Non-standard word order, non-standard use of conjunctions, cases, grammatical number, high number of demonstrative pronouns and certain particles.
- In the 49 % of the linguistically and technically non-standard tweets.
- These phenomena do not pose a problem for the annotation, however need to be linguistically addressed in the future.
- What is non-standard? – Mind the gap: no corpus-based description of contemporary Slovene.

Syntactical Standardisation

Analysis of the dataset, establishing inter-annotation agreement for:

1. Non-standard word order.

   Moram se danes zvečer dol usesti in urediti zadeve s komentarjem za Fejs.

   Danes zvečer se moram usesti dol in urediti zadeve s komentarjem za Fejs.

   (I have to sit down tonight and arrange the things regarding the Facebook comment.)

2. Non-standard use of conjunctions.

   jah saj za manj recimo tudi jaz ne bi peljal **samo** dobro oni jih malo več peljejo

   Jah, saj za manj recimo tudi jaz ne bi peljal. **Ampak** dobro, oni jih peljejo malo več.

   (Well, I wouldn‘t drive for less eather just ok they drive more of them.)

Syntactical Standardisation

Analysis of the dataset, establishing inter-annotation agreement for:

3. Erroneous use of cases, grammatical number, and similiar categories.

   ne vem **sestavine,** ker dobro prenaša. mogoče so naravne.

   Ne vem **sestavin,** ker dobro prenaša. Mogoče so naravne.

   (I don‘t know the ingredients because she tolerates well. They might be natural.)

4. Non-standard fillers.

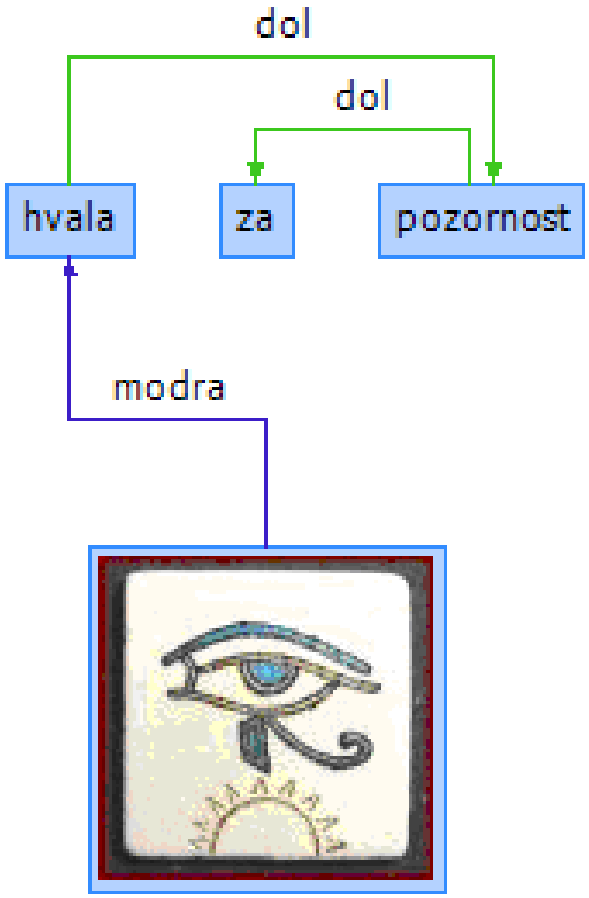
   Najboljši je pa Caleb ko **ga** tam vmes pleše.

   Najboljši pa je Caleb, ko tam vmes pleše.

   (And Caleb is the best dancing it there.)

Conclusion

- The JOS dependency model in combination with the SSJ annotation program proved to be adequate for the described task.
- Pros: robustness; multiple attachments to the root element.
- Cons: little possibility for cross-lingual comparison. -> Universal Dependencies (Dobrovoljc et al., 2016).
- Further annotation of language data (from Twitter or other CMC genres); the training of a parser for the selected CMC domain(s); qualitative linguistic analyses of the annotated dataset.


dol
dol
hvala
za
pozornost
modra

References

Arhar Holdt, Š. and Dobrovoljc, K. (2015). Zveze samostalnika z nesklonljivim levim prilastkom v korpusih Janes in Kres. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 4–9.

Böhmová, A., Hajič, J., Hajičová, E. and Hladká, B. (2003). The Prague dependency treebank. In *Treebank: Building and Using Parsed Corpora*. Netherlands: Springer, pp. 103–127.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J. and Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 29(2), pp.1-30.

Čibej, J., Fišer, D. and Erjavec, T. (2016a). Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets. In *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož: ELRA, pp. 5–10.

Čibej, J., Arhar Holdt, Š., Erjavec, T. and Fišer, D. (2016b). Razvoj učne množice za izboljšano označevanje spletnih besedil. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Crystal, D. (2011). *Internet Linguistics: A Student Guide.* London, New York: Routledge.

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '16).* Portorož, pp. 1566–73.

Dobrovoljc, K., Erjavec, T. and Krek, S. (2016). Pretvorba korpusa ssj500k v Univerzalno odvisnostno drevesnico za slovenščino. In *Proceedings of the Conference on Language Technologies and Digital Humanities.* Ljubljana (in print).

Dobrovoljc, K., Krek, S. and Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. In T. Erjavec, J. Žganec Gros (Eds.), *Zbornik Osme konference Jezikovne tehnologije.* Ljubljana: Institut Jožef Stefan, pp. 42–47.

Erjavec, T., Fišer, D., Krek, S. and Ledinek, N. (2010). The JOS linguistically tagged corpus of Slovene. In: *LREC 2010, 7th International Conference on Language Resources and Evaluations*. Valletta, pp. 1806–1809.

Erjavec, T., Ignat, C., Pouliquen, B. and Steinberger, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference*. Poznan, pp. 32–36.

Fišer, D., Ljubešić, N. and Erjavec, T. (2015). The JANES corpus of Slovene user generated content: construction and annotation. In *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts*. Rennes,.p. 11.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S. and Velušček, A. (2008). *Specifikacije za učni korpus.* Kamnik: Projekt »Sporazumevanje v slovenskem jeziku« ESS in MŠŠ.

Jakop, N. (2008). Pravopis in spletni forumi – kva dogaja? In *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*, pp. 315–327.

Kranjc, A. and Robnik Šikonja, M. (2015). Postavljanje vejic v slovenščini s pomočjo strojnega učenja in izboljšanega korpusa Šolar. In D. Fišer (Ed.), *Zbornik konference Slovenščina na spletu in v novih medijih,* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 38–43.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C. and Smith, N. A. (2014). A dependency parser for tweets. In *Proc. of EMNLP*. Doha, Qatar, pp. 1001–1012.

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N. and Holz, N. (2015). *Training corpus ssj500k 1.4, Slovenian language resource repository CLARIN.SI*, http://hdl.handle.net/11356/1052.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference.* Hissar, pp. 371–378.

Ljubešić, N., Erjavec, T. and Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal: Proceedings: part II.* Springer, Heidelberg, pp. 164–175.

Michelizza, M. (2015). *Spletna besedila in jezik na spletu.* Založba ZRC, ZRC SAZU, Ljubljana.

Myslin, M. and Gries, S. T. (2010). k dixez? A corpus study of Spanish Internet orthography. *Literacy and Linguistic Computing*, 25(1), pp. 85–104.

Storrer, A. (2013). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*. De Gruyter Mouton, pp. 171–196.

Zwitter Vitez, A. and Fišer, D. (2015). From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: proceedings of eLex 2015 Conference*, *Herstmonceux Castle, UK*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Birmingham: Lexical Computing, pp. 250–267.