2018-09-20

hr500k – A Reference Training Corpus of Croatian.

Nikola Ljubešić
Jožef Stefan Institute, nikola.ljubesic@ijs.si

Željko Agić
University of Copenhagen, zeag@itu.dk

Filip Klubicka
Technological University Dublin, d17124386@mydit.ie

See next page for additional authors

Recommended Citation

Ljubešić, N., Agić, Z. & Klubicka, F. (2018). hr500k – A reference training corpus of Croatian, *Language Technologies and Digital Humanities Conference*, Ljubljana, Slovenia, 20-21 September. doi:10.21427/0pjb-f168

hr500k – A Reference Training Corpus of Croatian

Nikola Ljubešić,[*] Željko Agić,[†] Filip Klubička,[‡] Vuk Batanović,[§] Tomaž Erjavec[*]

[*]Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si, tomaz.erjavec@ijs.si

[†]Department of Computer Science, IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
zeag@itu.dk

[‡]ADAPT Centre, School of Computing, Dublin Institute of Technology,
Kevin Street, Dublin, Ireland
filip.klubicka@adaptcentre.ie

[§]School of Electrical Engineering, University of Belgrade
Innovation Center, School of Electrical Engineering, University of Belgrade
Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
vuk.batanovic@ic.etf.bg.ac.rs

Abstract

In this paper we present hr500k, a Croatian reference training corpus of 500 thousand tokens, segmented at document, sentence and word level, and annotated for morphosyntax, lemmas, dependency syntax, named entities, and semantic roles. We present each annotation layer via basic label statistics and describe the final encoding of the resource in CoNLL and TEI formats. We also give a description of the rather turbulent history of the resource and give insights into the topic and genre distribution in the corpus. Finally, we discuss further enrichments of the corpus with additional layers, which are already underway.

1. Introduction

Natural language processing techniques today are primarily based on supervised machine learning. Reference training corpora are therefore crucial for the development of NLP tools such as taggers, parsers, named entity recognizers etc.

In this paper we present hr500k – a reference corpus of Croatian which is currently annotated on the following levels: (1) token, sentence, and document segmentation, (2) morphosyntax, (3) lemmas, (4) dependency syntax, (5) semantic roles and (6) named entities. The corpus presents a significant extension of previous training corpora developed for Croatian, namely the SETimes.HR and the SETimes.HR+ corpora, and allows Croatian's basic language technologies to finally catch up to other well-equipped Slavic languages such as Slovene (Krek et al., 2018), Czech (Hajič et al., 2012) or Polish (Broda et al., 2012).

2. Description of the corpus

The hr500k corpus consists of 900 documents segmented into around 25 thousand sentences, making the average document length around 28 sentences or 563 tokens, while the average sentence length is around 20 tokens. A statistical overview of the corpus is given in Table 1. Each document is preceded by a tag indicating its name and the URL of the source, if available. Each tokenized sentence is preceded by a tag stating its original, untokenized text. The form of all such tags is compliant with the Universal Dependencies v2 specifications.[1]

Item	Count
Documents	900
Sentences	24 794
Tokens	506 457
Types	73 548
Lemmas	34 329
MSDs	768

Table 1: A statistical overview of the hr500k corpus

Regarding the genres covered in the corpus, around 55% of the content are news articles, blogs covering 20% of the content, forums 15% and 10% being covered by other genres. For the topical distribution, around 50% of content covers the general topic, music and medicine each covering around 10% of the content, while business, tech, lifestyle and education cover around 5% of the content each. A more detailed description of the genre and topic distributions, from the perspective of extending the corpus through time, can be found in Section 4..

2.1. Morphosyntax and lemmas

The entire hr500k corpus is annotated with morphosyntactic tags and lemmas for each token. Morphosyntax is encoded according to the MULTEXT-East V5 guidelines,[2] which specify 13 part-of-speech categories, with numer-

[1] http://universaldependencies.org/v2/
[2] http://nl.ijs.si/ME/V5/msd/html/

MTEv5 gloss	POS tag	Count	Percentage
Nouns	N	135 822	26.82%
Verbs	V	80 499	15.89%
Punctuation	Z	62 116	12.26%
Adjectives	A	50 982	10.07%
Adpositions	S	45 145	8.91%
Pronouns	P	40 591	8.01%
Conjunctions	C	36 685	7.24%
Adverbs	R	26 051	5.14%
Numerals	M	12 744	2.52%
Particles	Q	8 787	1.74%
Residuals	X	5 051	1.00%
Abbreviations	Y	1 666	0.33%
Interjections	I	318	0.06%

Table 2: MTEv5 part-of-speech tag distribution in the hr500k corpus

ous morphosyntactic attributes particular to each category. A list of these categories, alongside their frequencies in the hr500k corpus, is given in Table 2. In addition to the MTEv5 specification, we also provide POS tags in accordance with the Universal Dependencies v2 standard, which describes 17 part-of-speech categories. The frequency distribution of UD POS tags in the hr500k corpus is shown in Table 3.

MTE morphosyntactic tags can, for the most part, be automatically mapped into UD POS tags, and we provide the mapping table and code on the hr500k Github repository.[3] The only exception to this are abbreviations (MULTEXT-East tag `Y`), which have to be converted manually.

Given that the corpus was extended through time (more on the history of the corpus will be reported in Section 4.), there were multiple annotation rounds on the morphosyntactic and lemma annotation layers. To secure a consistent annotation of these phenomena, we have recently performed a series of global annotation consolidation procedures by which we expect for the annotation consistency to be high. However, up to this point, we have not measured this level of consistency.

2.2. Dependency syntax

The first two fifths of the hr500k, i.e. the first 197 028 tokens of the corpus, have been annotated with regard to dependency syntax. These annotations include both the older syntactic tags presented by Agić and Ljubešić (2014), as well as the newer Universal Dependency v2 syntactic relations.[4] Table 4 shows the distribution of the UDv2 syntactic tags in our corpus.

This annotation layer was annotated by a single annotator and there were no annotation consolidation procedures performed up to this point. One of our future goals is to harmonize the UD annotations between the Slovene (Krek et al., 2018), Croatian and Serbian (Samardžić et al., 2017) training corpora.

[3] `http://github.com/nljubesi/hr500k/`

[4] The meaning of the tags is explained here: `http://universaldependencies.org/u/dep/index.html`

UD POS gloss	UD POS tag	Count	Percentage
Nouns	NOUN	113 674	22.44%
Punctuation	PUNCT	61 914	12.22%
Adjectives	ADJ	56 071	11.07%
Verbs	VERB	49 089	9.69%
Adpositions	ADP	45 144	8.91%
Auxiliary	AUX	31 413	6.20%
Adverbs	ADV	26 144	5.16%
Proper nouns	PROPN	23 160	4.57%
Coord. conj.	CCONJ	22 175	4.38%
Determiners	DET	21 012	4.15%
Pronouns	PRON	19 579	3.86%
Subord. conj.	SCONJ	14 510	2.86%
Particles	PART	8 941	1.76%
Numerals	NUM	7 813	1.54%
Other	X	5 262	1.04%
Interjections	INTJ	322	0.06%
Symbols	SYM	234	0.05%

Table 3: UD part-of-speech tag distribution in the hr500k corpus

2.3. Semantic roles

Semantic roles are currently annotated in the oldest part of the corpus, namely the documents coming from the original SETimes.HR corpus, without the original testing portion of the corpus. This part of the corpus contains 163 documents, 3 757 sentences and 83 630 tokens. On average there are 5 SRL labels applied to each sentence.

The SRL formalism was developed inside a bilateral Slovene-Croatian project, lead by Simon Krek on the Slovene side and Kristina Štrkalj Despot on the Croatian side. The formalism presents for the most part a simplification of the Prague Dependency Treebank formalism. Table 5 shows the distribution of the SRL tags in our corpus.

The Croatian SRL annotations were applied by a single annotator, with complex examples being discussed together by the Croatian and Slovene project partners.

2.4. Named entities

Named entity annotations cover the entire hr500k and are encoded in the IOB2 format. 36 735 tokens, or 7.25% of the total, are marked as belonging to a named entity, of which there are 23 186, which means there are around 26 named entities per document, or almost one per sentence, on average. Five NE types are considered – the standard categories for people (PER), locations (LOC), organizations (ORG), and miscellaneous entities (MISC) are augmented with a person derivative category (DERIV-PER), intended for marking personal (possessive) adjectives, enabling better information extraction and anonymization of personal data. The annotation guidelines applied are those that were developed while annotating the Slovene ssj500k and Janes-Tag datasets.[5]

The distribution of entities in hr500k between these categories is given in Table 6. The distribution of tokens belonging to a named entity is given in Table 7.

[5] `http://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf`

UD syntactic tag	Count	Percentage
punct	23 894	12.13%
case	18 936	9.61%
nmod	18 289	9.28%
amod	18 229	9.25%
nsubj	13 959	7.08%
obl	12 226	6.20%
conj	9 414	4.78%
root	8 889	4.51%
obj	8 719	4.42%
aux	8 532	4.33%
advmod	8 103	4.11%
cc	7 528	3.82%
mark	3 941	2.00%
acl	3 817	1.94%
det	3 518	1.78%
cop	3 478	1.76%
xcomp	2 949	1.50%
appos	2 894	1.47%
nummod	2 887	1.46%
flat	2 526	1.28%
compound	2 274	1.15%
parataxis	2 268	1.15%
expl	2 173	1.10%
discourse	2 040	1.04%
ccomp	1 766	0.90%
advcl	1 753	0.89%
fixed	707	0.36%
iobj	689	0.35%
csubj	359	0.18%
orphan	152	0.08%
goeswith	55	0.03%
list	24	0.01%
vocative	23	0.01%
dep	9	0.01% <
dislocated	8	0.01% <

Table 4: UD syntactic relation distribution in the hr500k corpus

SRL gloss	SRL tag	Count	Percentage
Patient	PAT	4 860	26.10%
Agent	ACT	4 731	25.40%
Result	RESLT	2 860	15.36%
Time	TIME	1 344	7.22%
Recipient	REC	603	3.24%
Modality	MODAL	586	3.15%
Location	LOC	525	2.82%
Manner	MANN	472	2.53%
Duration	DUR	351	1.88%
Origin	ORIG	268	1.44%
Cause	CAUSE	242	1.30%
Aim	AIM	224	1.20%
Regard	REG	222	1.19%
Goal	GOAL	202	1.08%
Event	EVENT	170	0.91%
Means	MEANS	169	0.91%
Quantity	QUANT	158	0.85%
MW predicate	MWPRED	134	0.72%
Accompaniment	ACMP	106	0.57%
Condition	COND	87	0.47%
Contradiction	CONTR	82	0.44%
Frequency	FREQ	81	0.43%
Part of phraseme	PHRAS	74	0.40%
Source	SOURCE	50	0.27%
Restriction	RESTR	22	0.12%
Total		18 623	100%

Table 5: Distribution of SRL tags in the hr500k corpus

Named entity type	Count	Percentage
Person	6 802	29.34%
Person derivative	317	1.37%
Location	6 214	26.80%
Organization	6 354	27.40%
Miscellaneous	3 499	15.09%
Total	23 186	100%

Table 6: Distribution of named entities in the hr500k corpus

The named entity annotation layer was applied by two annotators, with collisions in the annotations being resolved by a super-annotator. Regardless of the double-annotation procedure, we plan to perform a global lexical label consolidation procedure in the near future.

3. Corpus encoding and publishing

The working version of the corpus was encoded in a modified version of the tabular CoNLL-X format (Buchholz and Marsi, 2006), consisting of the following columns:

1. ID: sentence-local word index
2. FORM: token, i.e. word form or punctuation symbol
3. LEMMA: lemma of word form
4. POS: part-of-speech according to the MULTEXT-East specifications
5. MSD: morphosyntactic description according to the MULTEXT-East specifications
6. MSDFEAT: morphosyntactic features according to the MULTEXT-East specifications
7. SETDEPREL: dependency relation (head, label) according to the SETimes formalism
8. UDDEPREL: dependency relation (head, label) according to the UDv2 formalism

Named entity type	Token count	Percentage
Person	10 241	2.02%
Person derivative	319	0.06%
Location	7 445	1.47%
Organization	11 216	2.21%
Miscellaneous	7 514	1.48%
Total	36 735	7.25%

Table 7: Distribution of named entity tokens with regard to the whole hr500k corpus

9. UPOS+FEATS: universal part-of-speech tag with features from the universal feature inventory
10. UDSPEC: UDv2 language-specific features
11. NER: named entity annotations encoded through IOB2
12. SRLHEAD: heads of semantic roles, order of occurrence in a sentence defines in which columns semantic roles to specific heads are encoded (columns 13-23)
13. SRL: semantic roles, encoded to column 23

The CoNLL-type format was converted to TEI, i.e., to a schema following the Guidelines for Electronic Text Encoding and Interchange (TEI Consortium, 2017) to ensure (meta-)data persistence. Apart from the automatic conversion of the text and its annotations, this also involved writing the `teiHeader` element, which gives the meta-data of the corpus, containing its name, authors, license, source description, annotation vocabulary, tag usage, revision history etc.

As illustrated in Figure 1, each sentence in the TEI encoding is assigned a unique ID, as are all the tokens (words, `w` and punctuation symbols, `pc` in the sentence; white space in the sentence is also marked-up with `c`).

The lemma of the words is given in the `@lemma` attribute, while all tokens are given their MULTEXT-East MSD in the `@ana` attribute. The UD parts of speech and features are given in the `@msd` attribute, which is an attribute newly introduced into the TEI. Note that the double pipe symbol is used to separate the universal features from the (Croatian) language-specific ones. The reason why the MULTEXT-East MSDs are not given in the `@msd` attribute, as might be expected, is that while `@msd` can contain any string, the `@ana` is defined as a pointer, which MULTEXT-East MSDs can be, but UD features cannot; we explain below in more detail the functioning of TEI pointers for linguistic labels as used in the hr500k corpus.

Named entities are also encoded in-line, simply using the standard TEI `name` element, with the `@type` giving the type of the name.

The final three layers of annotation, namely the original syntactic dependencies, the UD dependencies and the semantic role labels are all encoded in stand-off annotation, using the `linkGrp` (link group) element, which specifies its type (so, annotation layer), the ordering of the arguments of the links, and then contains the links themselves; each of these gives the link label and pointers to the IDs of the link head and argument. It should be noted that in cases where a syntactic dependency has the (virtual) root as its head, the references to the sentence ID is given (so, in the example above that would be `#train-s2`).

As mentioned, the `@ana` attribute is a pointer, which usually contains a local reference to an ID (e.g. `#train-s2.1`) or a fully qualified URI. TEI has another option for its pointers, namely using a prefix before the ID and separated from it by a colon (e.g. `mte:Npnsn`). Such pointers are then resolved using the `prefixDef` element in the TEI header, which defines the prefixing schema used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the hr500k corpus all the prefixes are simply expanded to local references, which are given in the TEI header, except for the MULTEXT-East MSD, which are defined in the `back` element of the TEI document. There, each MSD is defined as a feature-structure giving the decomposition of the MSD into its features. It is thus a simple matter, using just the TEI encoded corpus, to move from `mte:Mdo` to `Category = Numeral, Form = digit, Type = ordinal`.

The TEI encoded corpus, which is to be taken as the canonical version of the hr500k corpus, was then automatically converted to the so called vertical format which is used by CQP-based concordancers, in particular (no)Sketch Engine (Rychlý, 2007). The vertical format is able do encode hierarchical structures (e.g. sentences and names), and token annotations (e.g. lemmas and MSDs), but not links between tokens (e.g. dependencies and semantic role labels). To nevertheless preserve as much of this information as possible, the dependencies are annotated next to tokens, so that the argument token is annotated with the dependency label and head lemma.

Finally, the TEI, vertical and CoNLL encoded corpus were deposited to the CLARIN.SI repository,[6] where the corpus is available under the CC BY-SA license. The corpus is also available for exploration under the CLARIN.SI noSketch Engine and KonText concordancers; the links are given on the CLARIN.SI repository landing page.

4. History of the corpus

4.1. The SETimes.HR corpus

The prolific five-year period between the seminal shared tasks in dependency parsing (Buchholz and Marsi, 2006; Nivre et al., 2007) and the emergence of the first cross-linguistically uniform morphological and syntactic annotation guidelines (De Marneffe and Manning, 2008; Petrov et al., 2011; McDonald et al., 2013) has thoroughly changed the landscape of multilingual NLP.

Back then, the field's positive momentum was not entirely mirrored by the developments for Croatian. In 2007 the Croatian dependency treebank (HOBS), deemed a central resource for training basic NLP models, was but a 50-sentence prototype (Tadić, 2007). By late 2012 the resource grew to around 75% of its full size of 4.6k sentences (Berović et al., 2012). Yet, it was not publicly available. At that point in time there were *no* freely available Croatian language resources to train NLP models whatsoever.

The consequences were dire. To illustrate, in 2012 one could not even tag Croatian for POS, let alone perform any syntactic parsing, while the rest of the field was involved in pursuing human-level accuracies in these basic tasks. The SETimes.HR corpus was thus developed to address the urgent need for a free-culture resource to build and evaluate basic NLP for Croatian.

The original instantiations of SETimes.HR are the experiments in POS tagging and dependency parsing by Agić et al. (2013a; 2013b), which also included a cross-lingual application to Serbian as a very closely related yet truly low-resource language. Agić and Ljubešić (2014) provide

[6] published at `http://hdl.handle.net/11356/1183`

```
<s xml:id="train-s2">
  <name type="loc">
    <w xml:id="train-s2.1" lemma="Kosovo" ana="mte:Npnsn"
       msd="UposTag=PROPN|Case=Nom|Gender=Neut|Number=Sing">Kosovo</w>
  </name>
  <c> </c>
  ...
  <w xml:id="train-s2.9" lemma="pritužba" ana="mte:Ncfpg"
     msd="UposTag=NOUN|Case=Gen|Gender=Fem|Number=Plur||SpaceAfter=No">pritužbi</w>
  <pc xml:id="train-s2.10" ana="mte:Z" msd="UposTag=PUNCT">.</pc>
  <linkGrp targFunc="head argument" type="SE-SYN">
    <link ana="se-syn:Sb" target="#train-s2.3 #train-s2.1"/>
    <link ana="se-syn:Adv" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
  <linkGrp targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:nsubj" target="#train-s2.3 #train-s2.1"/>
    <link ana="ud-syn:advmod" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
  <linkGrp targFunc="head argument" type="SRL">
    <link ana="srl:ACT" target="#train-s2.3 #train-s2.1"/>
    <link ana="srl:MANN" target="#train-s2.3 #train-s2.2"/>
    ...
  </linkGrp>
</s>
```

Figure 1: TEI encoding of a corpus sentence

a thorough documentation of the new corpus and introduce an annotation layer for named entities following Tjong Kim Sang and De Meulder (2003). The release was open to the public for all uses.

The corpus contained around 4,000 sentences (84k word forms) annotated using a modified MULTEXT-East V5 guideline for POS and morphology, while the syntactic dependencies followed a novel Prague-motivated lean tagset by Agić and Merkler (2013). It was introduced to the first version of Universal Dependencies (Agić et al., 2015; Nivre et al., 2016) together with compliant annotations that were applied manually.

As a grassroots effort, the SETimes.HR corpus was not devoid of flaws. For one, its training section consisted entirely of newspaper data from a single source,[7] while its test data were multi-domain. However, it played a crucial role in democratizing Croatian NLP resources, eventually drawing out even the dormant HOBS after years of public unavailability (Agić et al., 2014).

This section of the corpus is identifiable in the final hr500k corpus by document IDs starting with `set.hr` and consists of 164 documents (163 training and one testing document).

4.2. The SETimes.HR+ corpus

Further extensions of the SETimes.HR corpus were performed in 2014, in two main phases. The first phase consisted of adding texts collected in 2012 for a named entity recognition task. The second phase focused on selecting and annotating texts in a crowdsourcing framework inside a master course.

The first extension of the SETimes.HR corpus consisted of texts from a named entity recognition training corpus for Croatian that was built in 2012 during a student project and contained initially 59,212 tokens (Ljubešić et al., 2012). The data came from four different web domains belonging to the genres of general news, ICT news and business news. This section of the corpus is identifiable in the final hr500k corpus by document IDs starting with `news.hr`, consisting overall of 83 documents.

The second extension of the SETimes.HR corpus was performed as part of a master course. No specific topic domain was chosen, but rather a random sample of sentences from the general web which, through crowdsourcing efforts, were deemed as being of an acceptable linguistic standard. This dataset of 50,322 tokens was then automatically MSD-tagged, followed by employing crowdsourcing and a small team of experts to correct the annotations of tokens that were tagged differently by a tagger ensemble (Klubička and Ljubešić, 2014). This section of the corpus is organized in the final hr500k corpus into a single document with the document ID `web.hr`.

Both these corpus sections were later merged with the original SETimes.HR corpus into one corpus, internally referred to as SETimes.HR+, with approximately 190 thousand tokens in size. This new corpus was manually inspected for possible errors and inconsistencies. As for genre and register, the content of this corpus belonged mainly to news (>85%), and a little bit of the general web, which varied greatly by genre and topic, including the odd forum discussion or blog post, but mostly consisting of re-

[7] The now defunct Southeast European Times online news portal `setimes.com`, also basis of the SETimes parallel corpus.

	articles	blogs	forums	other
320k extension	40%	30%	20%	10%
hr500k	57.63%	20.6%	14.64%	7.13%

Table 8: Token percentage per web genre in the 320k extension added to the hr500k corpus and the final hr500k corpus

ports on politics, sports, religion, in addition to news and other informative articles.

The SETimes.HR+ corpus as described in this section currently also serves as the Croatian part of the current Universal Dependencies treebanks release (v2.2). It was manually annotated for UD-style syntactic dependencies, and its POS tags and morphological features were semi-automatically converted to UD.

4.3. hr500k

In 2015 we raised the bar to 500 thousand tokens (Ljubešić et al., 2016), motivated by results on morphosyntactic annotation of Slovene (Ljubešić and Erjavec, 2016) which showed that corpus supervision has a much higher impact on tagging accuracy than lexicon supervision. Thus, for the final phase of corpus assembly we manually selected 320k tokens worth of suitable documents from the hrWaC web corpus (Ljubešić and Klubička, 2014). The documents were automatically morphosyntactically annotated with a tagger learned on the 190k-sized SETimes.HR+ corpus, which was followed by having experts perform manual correction of the automatic annotations.

However, this time around we were somewhat spoiled for choice with regards to the content to be included in the corpus, as the hrWaC corpus boasts 1.3 billion tokens. This allowed us the luxury to include more varied content than the previous iterations had. Thus, our aim was to gather a representative sample of the Croatian language; one that expands beyond the confines of a particular genre, topic or register, and includes many different examples of linguistic expression that can be found on the web. With accordance to that, we divided the additional 320k token sample into 4 sections according to web genre, in the ratios shown in the first row of Table 8, the second row showing the distribution of web genres in the final hr500k resource.

This way, we covered registers used in different kinds of genres – articles, blogs, forums, reviews and advertisements – while at the same time covering a wide range of topics that were inadequately or not at all covered in the SETimes.HR+ corpus, which mainly consisted of general news articles. The web domains that we included cover topics ranging from medicine, education and technology, through music, sports and religion, all the way to listings and adverts, literature and political activism. Where possible, we also made the effort to include any user comments posted on their corresponding articles and blogs, so that, coupled with forum discussions, the corpus would also include a sample of the language used in direct communication among Internet users. Such meticulous selection results in considerable variety among documents, but given that documents were selected exclusively from a list of the top 200 most frequent domains in the hrWaC corpus, this varied sample is actually quite representative of the Croatian web.

topic	token ratio	topic	token ratio
general	35.01%	business	4.41%
music	13.55%	listings	3.88%
medicine	12.26%	religion	3.81%
tech	7.93%	sports	3.59%
lifestyle	7.38%	culture	2.36%
education	5.80%		

Table 9: Topic domain distribution in the 320k extension

topic	token ratio	topic	token ratio
general	51.89%	education	3.61%
music	8.43%	religion	2.87%
medicine	7.63%	sports	2.74%
business	6.93%	listings	2.42%
tech	6.92%	culture	1.97%
lifestyle	4.59%		

Table 10: Topic domain distribution in the hr500k corpus

An approximation of the distribution of web genres in the final hr500k corpus created by merging all the hitherto described corpora is presented in the second row of Table 8. An overview of the topic domains that enriched the corpus in the second phase of construction is presented in Table 9 and is based on the general topic of the web domains the sentences come from, while an approximation of topic domain distribution in the final 500k corpus is presented in Table 10. Compared to the approximate >85% of general news articles that comprise the initial SETimes.HR+ corpus, this is a vast improvement in terms of data diversity.

5. Conclusion

In this paper we presented the manually annotated reference corpus of Croatian, which is currently the largest and richest training dataset for Croatian and is made freely (CC BY-SA) available for download[8] and for on-line exploration.[9]

Future plans are primarily directed at (1) further consolidation of the presented annotation layers and (2) extension to new annotation layers, two being planned for the near future in form of layers of verbal multiword expression annotations.

We will also define training, development and test portions of the corpus for each task and benchmark the available language tools for the available tasks, thereby further fostering development of language technologies for Croatian and other languages.

6. Acknowledgements

The presented resource was initially developed through pure enthusiasm of a network of researchers, with recent extensions and consolidations being funded by the Regional Linguistic Data Initiative (ReLDI, Swiss National

[8] http://hdl.handle.net/11356/1183

[9] https://www.clarin.si/kontext/first_form?corpname=hr500k

Science Foundation grant IZ74Z0 160501) and the Slovenian Research Infrastructure CLARIN.SI. Filip Klubička is also supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Vuk Batanović is also partially supported by the III44009 research grant of the Ministry of Education, Science and Technological Development of the Republic of Serbia.

7. References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Danijela Merkler. 2013. Three syntactic formalisms for data-driven dependency parsing of croatian. In *International Conference on Text, Speech and Dialogue*, pages 560–567. Springer.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013a. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the Fourth Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Željko Agić, Danijela Merkler, and Daša Berović. 2013b. Parsing Croatian and Serbian by using Croatian dependency treebanks. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*.

Željko Agić, Daša Berović, Danijela Merkler, and Marko Tadić. 2014. Croatian dependency treebank 2.0: New annotation guidelines for improved parsing. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*.

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, et al. 2015. Universal dependencies 1.1. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague*, 3.

Daša Berović, Željko Agić, and Marko Tadić. 2012. Croatian dependency treebank: Recent development and initial experiments. In *Seventh International Conference on Language Resources and Evaluation (LREC 2012)*.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. Kpwr: Towards a free corpus of polish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*.

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. *Prague Czech-English Dependency Treebank 2.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11858/00-097C-0000-0015-8DAF-4`.

Filip Klubička and Nikola Ljubešić. 2014. Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of croatian. In Tomaž Erjavec and Jerneja Žganec Gros, editors, *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*, Ljubljana, Slovenia.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, and Anja Zajc. 2018. *Training corpus ssj500k 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1181`.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Marija Stupar, and Tereza Jurić. 2012. Building named entity recognition models for croatian and slovene. In Tomaž Erjavec and Jerneja Žganec Gros, editors, *Proceedings of the Eighth LANGUAGE TECHNOLOGIES Conference*, Ljubljana, Slovenia.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Nikola Ljubešić and Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: The case of slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald,

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal dependencies for serbian in comparison with croatian and other slavic languages. In *The 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*.

Marko Tadić. 2007. Building the croatian dependency treebank: the initial stages. *Suvremena lingvistika*, 63(1):85–92.

TEI Consortium. 2017. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.