Croatian Error-Annotated Corpus of Non-Professional Written Language

Vanja Štefanec*, Nikola Ljubešić†, Jelena Kuvač Kraljević‡
* Center for Postgraduate Studies, University of Zagreb
Zvonimirova 8, HR-10000 Zagreb, Croatia
vanja.stefanec@gmail.com
† Dept. of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia
nljubesi@ffzg.hr
‡ Laboratory for Psycholinguistic Research, Department of Speech and Language Pathology, University of Zagreb
Zvonimirova 8, HR-10000 Zagreb, Croatia
jkuvac@erf.hr

Abstract

In the paper authors present the Croatian corpus of non-professional written language. Consisting of two subcorpora, i.e. the clinical subcorpus, consisting of written texts produced by speakers with various types of language disorders, and the healthy speakers subcorpus, as well as by the levels of its annotation, it offers an opportunity for different lines of research. The authors present the corpus structure, describe the sampling methodology, explain the levels of annotation, and give some very basic statistics. On the basis of data from the corpus, existing language technologies for Croatian are adapted in order to be implemented in a platform facilitating text production to speakers with language disorders. In this respect, several analyses of the corpus data and a basic evaluation of the developed technologies are presented.

Keywords: error corpus, language disorders, Croatian language

1. Introduction

In this paper we describe the Croatian corpus of non-professional written language (Kuvač Kraljević et al., in press) that is based on written language production of two groups of participants: healthy speakers and speakers with various types of language disorders (LDs). This resource, created in an interdisciplinary environment, offers valuable data for research in different fields of science. It provides data about language production which can be of interest to speech and language pathologists working in clinical settings, but also to linguists and neuroscientists studying language processing in general. Besides that, it is a unique language resource that can be used for improving existing or building new and specialized language technologies (LTs). Roughly containing 0.5M tokens, it is, as to our best knowledge, the biggest corpus of this type in general.

It is needless to say that resources of this type are scarce. Moreover, this is not the case only in linguistically under-resourced languages like Croatian. Specialized corpora which contain samples of written language produced by speakers with developmental or acquired language disorders, i.e. error corpora, are generally rare (with the exception of the learner corpora, which, we believe, capture different types of errors than those targeted here), and especially those entirely composed of samples produced by the speakers with language disorders. But, nevertheless, besides for scientific research of the underlying phenomena, such corpora are irreplaceable for developing various language tools, such as contextually-aware spelling correctors. In the absence of a better resource, needed corpora are sometimes artificially created by introducing errors into correct text (Pedler, 2007, p. 43). Other commonly used resources (and often wrongly referred to as 'corpora') are simply lists of commonly misspelled words, extracted from every context, paired with their correctly written equivalents. These are, for example, the Birkbeck Error Corpus (Mitton, 1985) or the Wikipedia list of common misspellings (Wikipedia, 2015). One of the rare examples of a resource that keeps the spelling errors in their context is more of a document collection than a corpus (Pedler, 2007). Although remarkable by its size of 12,000 tokens, it is composed of various available materials collected from various sources without a defined sampling methodology. Among the examples of full-fledged (yet very small by its size of 1k tokens) corpora, we could mention the work of Rello et al. (2012) who approached the problem of collecting the corpus of texts produced by adolescents with dyslexia with a solid methodology. Worth mentioning are also the attempts of automated building of error corpora by Miłkowski (2007), and those by Rodrigues and Rytting (2012), using crowdsourcing principles.

Our corpus has already served as the base for different streams of research. Kuvač Kraljević et al. (2016) have tackled the methodological questions concerning collecting and sampling of specialized oral and written adult speakers corpora. In the framework of cognitive model of writing Kuvač Kraljević and Kologranić Belić (2015) have analyzed the grammatical and orthographic features in written language production of adolescents with specific language impairment (SLI). Recent study based on this corpus was oriented towards text quality measured by various discourse elements, such as coherence and cohesion, in subjects with developmental language disorders only (Kuvač Kraljević et al., under revison). Our focus in this paper is put on the use of the data in developing specialized language technologies, primarily predictors of following words and context-aware spelling correctors.

In the paper we primarily focus on measuring the amount and level of error produced by healthy speakers and speakers with various language disorders, along with the statis-

tical significance of the measured differences. These measurements provide us with data crucial for developing efficient or adapting existing language technologies that can be useful for both beneficiaries, healthy speakers and those with language disorders.

2. Resource Development

The corpus was collected during the period of 8 months by the speech and language pathologists who were dealing primarily with language disorders in medical and educational institutions. In order to cover multiple genres of written language, various text genres were elicited. The text collection procedure, as well as the eliciting material adapted for different age groups of the participants, were prescribed in advance. Accordingly, participants were asked to write several different kinds of texts (e.g. essay, narrative, dictation, official letter). Every item in the material was designed to elicit a written response of a roughly predictable length and a certain formal complexity. Except for data collected in the clinical setting from the participants with language disorders, which constitute our clinical subcorpus, the same group of interrogators collected roughly the same amount of data from the healthy participants in a non-clinical setting using the same material and methodology. The primary purpose of the collected data was to tune existing language technologies to characteristics of language disorders. The optimized language technologies (i.e. a predictor of following words and a context-aware spelling corrector) were then to be used in a platform designed for facilitating text production to speakers with language disorders. Although the platform will, of course, facilitate text production in a digital form (i.e. text input), most of the responses collected from the participants were produced in a handwritten form and subsequently transcribed by the interrogators. Reasons for that are twofold. First, the age of participants ranged from 10 years to 80 years which leads to significant differences in their computer skills. And second, only those patients with traumatic brain injury or stroke who were in lesion phase (period from several weeks up to 5 months after onset) or late phase (period after 5 months, i.e. the rest of the patient's life) were encouraged to participate in written texts collection, still some of them were unable to type, and for some of them this kind of participation was very demanding. There were, however, several items in the material for which interrogators could, by their own discretion, give the participant the opportunity to produce his/hers response in a digital form. Text samples produced in a digital form can be identified in the corpus as they are marked as such and can be analyzed separately.

3. Resource Description

The corpus consists of 500 thousand tokens, out of which roughly 55% were produced by participants with language disorders. More than 36% of the tokens in this clinical subcorpus were produced by participants with dyslexia who are, along with participants with aphasia, the target population for the platform implementing LTs enhanced on the basis of the corpus.

Basic statistics about the entire corpus are given in Table 1.

4. Resource Annotation

The corpus was manually annotated by trained linguists on several linguistic levels. The first annotation layer consists of corrections of surface forms. Except for the corrections on the token level, annotators could merge multiple tokens into one and vice versa in case the participant incorrectly placed word boundaries or left out some syntactically non-optional element, e.g. a mandatory preposition. The second annotation layer contains error classifications into one or more of the 12 classes describing the scope of the errors. The scope of the errors could range from simple typos to semantic-related mistakes. Table 2 lists all the error classes used in the corpus.

The third and the fourth annotation layer consist of morphosyntactic annotations of both the original and the corrected surface forms. The reason we decided to perform this double annotation was to capture and explore possible systematic morphological and syntactical errors that could be related to some type of language disorder.

The morphosyntactic annotations used follow the revised Version 4 of the MULTEXT-East Morphosyntactic Specifications for Croatian (Ljubešić, 2013).

Annotators were instructed to intervene in the text as little as needed, and to correct only unintentional language and purely orthographic errors, while the use of non-standard language, regionalism or slang should be left as is. The purpose of such token normalization was to make the corpus as much as possible useful for the future development of text correction technologies. However, the decision on whether certain form should be corrected and classified as error was not always easy to make. For example, token $^{(*)}$*nebi* could be regarded both as a regional non-standard variant, or as an orthographically incorrect form of standard *ne bi* [⇒ "would not"], especially if compared with the regional form *nemrem* of standard *ne mogu* [⇒ "I can not"]. So, the guidelines given to the annotators were to normalize to standard only those forms which share the same phonological content with their standard equivalents.

Every text sample was annotated by only one annotator and no inter-annotator agreement was measured.

Table 3 gives the distribution of errors across language disorder statuses.

5. Statistical Analysis of the Resource

In this section we present a series of statistical analyses of the presented resource. We primarily focus on statistical descriptions that can help in developing LTs, concretely predictors of following words and spelling correctors. Although we are not aware of any relevant research that could back up this claim, our intuition tells us that there are certain non-negligible differences between different text input modalities, i.e. handwriting and typing, so we decided to analyze data that was collected through typing only. Also, since the text production-facilitating platform implementing these technologies will be offered to people with various types of language disorders, we treat developmental

[1]The numbers shown here differ slightly from those presented in Table 1, since one portion of the corpus containing the dictations is left unannotated.

	children ≤15 yrs.		young adults 16-20 yrs.		adults ≥21 yrs.		total	
	# of participants	# of tokens	# of participants	# of tokens	# of participants	# of tokens	# of participants	# of tokens
healthy speakers (HEALTHY)	17	22,648	16	29,871	101	192,881	134	245,400
language-related dyslexia (L-DYS)	76	72,771	4	6,954	7	11,609	87	91,334
visual-related dyslexia (V-DYS)	18	16,857	4	5,702	1	1,376	23	23,935
specific language impairment (SLI)	47	46,422	4	6,053	2	2,807	53	55,282
LDs related to intellectual disability (LD-ID)	6	6,398	0	0	0	0	6	6,398
LDs related to various syndromes (LD-VS)	2	2,004	0	0	0	0	2	2,004
dysgraphia (DYSG)	3	3,643	1	1,881	1	1,278	5	6,802
Broca's aphasia (BRO)	1	264	0	0	42	52,360	43	52,624
Wernicke's aphasia (WER)	0	0	0	0	1	860	1	860
anomic aphasia (ANOM)	0	0	0	0	7	11,386	7	11,386
other types of aphasia (APH)	0	0	0	0	14	19,515	14	19,515
traumatic brain injury (TBI)	0	0	7	8,613	19	24,412	26	33,025
total	170	171,007	36	59,074	195	318,484	388	548,565

Table 1: Basic corpus statistics

error class	example
typo	*popodme → popodne
orthography-related	*osijećam → osjećam
phonological (segment)	*Maia → Marija
phonological (syllable)	*poštoni → poštovani
morphological	*prijato → prijatelje
wrong inflection	*lavežem → lavežom
non-standard	*neda → ne da
neologism	*nordini → moždani
inflected neologism	*remu [*rema-ACC.sg] → remen
redundant	je je → je
syntactic	različitih → različitog
semantic	njezina → njegova

Table 2: Error classes

and acquired disorders as a unique clinical group. The presented corpus consists of 1891 such sentences produced by healthy subjects and 1034 sentences produced by subjects with language disorders.

The main questions we tackle in this section are:

1. do individuals with language disorders produce more spelling errors than healthy individuals?
2. do individuals with language disorders produce more spelling errors with a Damerau-Levenshtein distance (Damerau, 1964) higher than 2, i.e. spelling errors that would be very hard to correct with the traditional approach of identifying spelling corrections?
3. do individuals with language disorders introduce spelling errors earlier in a word, making thereby word predictors less useful?

In Table 4 we present the size of the two samples together with the number of misspellings. The results answer our first research question: subjects with language disorders do make almost 5 times more spelling errors than subjects with typical language processing. Given that our samples are not particularly big, we perform a chi-square test with the following null hypothesis: the number of correctly and incorrectly spelled tokens, and the clinical status of the subject are independent. As expected, the p-value of the test is $3 * 10^{-157}$, which enables us to safely reject our null hypothesis and conclude that individuals with various language disorders produce significantly ($p < .001$) more mistakes than healthy individuals.

Next, we investigate the distribution of the Damerau-Levenshtein distance among subjects of our two groups. The obtained results are presented in Table 5. Here we obtain the answer to our second question: do individuals with language disorders produce more spelling errors with a Damerau-Levenshtein distance higher than 2, which would make finding spelling corrections with the traditional approach almost impossible?[2] From the presented numbers it seems that this is not the case as the percentage of such misspellings is even slightly higher in the group of healthy subjects. Performing the chi-square test on the null hypothesis that the number of spelling errors with a Damerau-Levenshtein distance up to 2 and over two, and the clinical status of a subject are independent, we receive a p-value of 0.8783, because of which we can not reject the null hypothesis.

On the other hand, there is a visible difference between the percentage of spelling errors of distance 1 and 2. Therefore, we perform another chi-square test with the following null hypothesis: the number of spelling errors with distance 1 and distance 2, and the clinical status of subjects are independent. The test gives a p-value of 0.0324 which does enable us to reject the null hypothesis. Therefore we can conclude that participants with language disorders do make statistically significantly ($p < .05$) more spelling errors of

[2]Namely, the number of words that satisfy the criterion that the Damerau-Levenshtein distance to the misspelled word is 3 or higher becomes very large.

	# of annotated tokens[1]	typo	orthography	phon. (seg.)	phon. (syl.)	morphology	inflection	non-standard	neologism	neolog. (infl.)	redundant	syntactic	semantic
		$*10^{-3}$											
HEALTHY	203,711	1.84	3.15	3.01	0.23	0.38	0.72	2.59	0.15	0.02	0.42	0.65	0.27
L-DYS	75,451	4.64	21.17	36.57	2.49	2.65	3.41	8.91	1.29	0.15	1.27	4.48	1.64
V-DYS	18,775	5.11	18.22	26.90	1.92	2.13	3.25	7.94	1.44	0.11	0.59	2.72	1.49
SLI	45,858	2.46	13.80	23.66	1.40	2.27	2.88	5.34	0.52	0.09	0.89	3.45	1.18
LD-ID	5,155	3.30	15.32	22.70	3.10	2.52	6.40	6.60	0.58	0	1.16	4.46	0.97
LD-VS	1,632	0.61	44.73	64.95	4.90	12.25	12.25	23.28	0.61	0	0	11.64	4.29
DYSG	4,870	3.90	14.17	130.80	9.24	4.93	4.52	10.27	5.95	0.82	0.82	6.16	2.67
BRO	39,315	4.53	13.40	30.70	4.48	5.44	5.39	6.56	2.21	0.10	2.34	10.63	3.87
WER	623	1.61	22.47	17.66	3.21	1.61	1.61	3.21	0	0	1.61	3.21	1.61
ANOM	8,312	2.41	8.90	25.14	5.17	2.77	5.29	6.98	0.96	0.12	1.68	4.09	1.56
APH	14,170	3.74	14.82	23.36	4.94	4.16	4.16	7.55	0.92	0	1.27	6.35	2.12
TBI	23,651	4.27	13.53	18.94	2.49	2.16	3.13	10.02	1.27	0.17	1.94	3.17	1.56

Table 3: Probability of errors across language disorder statuses

	healthy	disorders
# of tokens	30,654	12,892
# of misspellings	342	695
% of misspellings	1.12%	5.39%

Table 4: Size of the analyzed datasets and the amount of misspellings

	healthy	disorders
DL=1	85.67%	81.58%
DL=2	7.60%	11.94%
DL>2	6.73%	6.00%

Table 5: Distribution of the Damerau-Levenshtein distance among healthy subjects and subjects with a language disorders

Damerau-Levenshtein distance 2 in comparison to spelling errors of distance 1 than healthy subjects.

From these results we can draw very useful conclusions for building a spelling corrector for participants with language disorders: while the number of spelling errors produced by subjects with language disorders is almost five times higher, the differences in the distribution of Damerau-Levenshtein distances among healthy subjects and those with language disorders up to distance of 2 and above 2 are not statistically significant. In both cases there is $\sim$6-7% of spelling errors that have a Damerau-Levenshtein distance to the correct form higher than 2, making the traditional approach of searching for spelling corrections not useful. One has, of course, bear in mind that among subjects with language disorders, these errors will still occur five times more frequently than among healthy participants, making their texts after spelling correction still less accurate. On the other hand, there is statistically significant difference in the number of spelling errors of distance 1 and of distance 2 between the two groups. However, both types of errors can be dealt with by using traditional spelling correction approaches.

character position	healthy	disorders
1	10.82%	13.24%
2	14.33%	14.39%
3	17.84%	16.55%
4	16.67%	17.84%

Table 6: Distribution of the position of the first misspelled character in misspelled words

We perform another set of analyses that are directed at the problem of predicting the following word while typing. First we present the distribution of the position of the first misspelled character in a word among healthy subjects and subjects with language disorders. The results on the first four characters, as later misspellings are not crucial for the technology in question, are given in Table 6. The two distributions seem quite similar, with the misspelling on the first character, which is the most dangerous one for word predictors, being $\sim 2\%$ more probable among participants with language disorders.

Given that there is an observable difference in the percentage of misspellings occurring on the first position in the two groups, we ran another chi-square test with the following null hypothesis: the number of misspellings on the first position in a word and later positions, and the language disorder status of a person are independent. The test gives a p-value of 0.3127, not allowing us to reject the null hypothesis.

After the second set of analyses we can conclude that word predictors should be, at least regarding the position of the first spelling error inside a word, as useful for individuals with language disorders as they are for healthy participants.

6. Evaluation of Language Technologies

In this section we present a final usage of the produced dataset, namely for evaluating language technologies for predicting the following word and spelling correction.

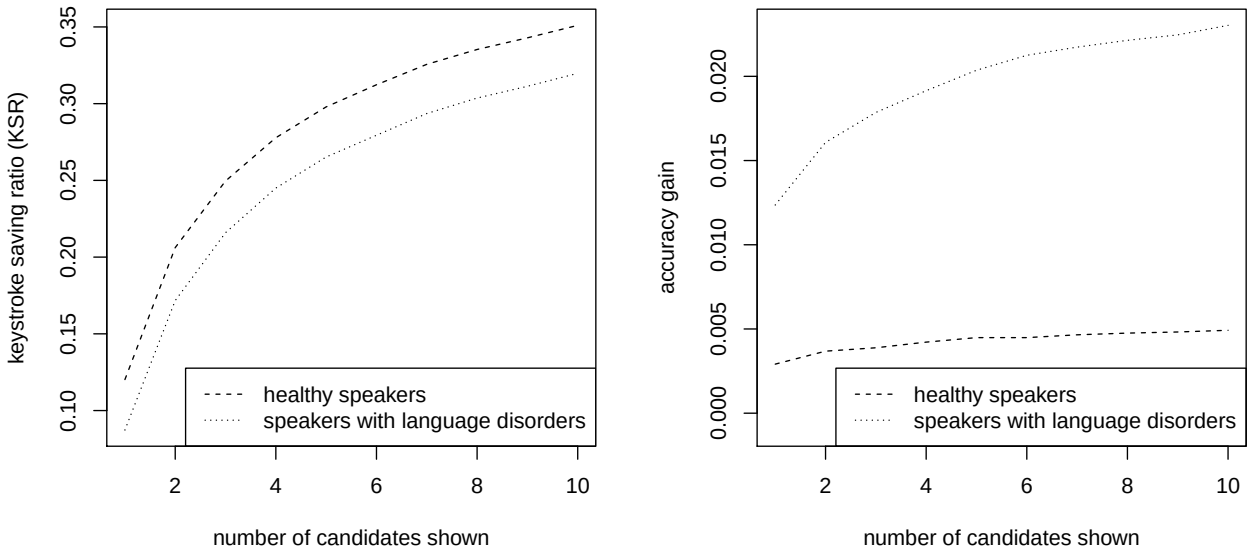

Figure 1: Evaluation results of the predictor of following words

6.1. Predictor of the Following Word

Our predictor of the following word is based on a character-level language model, built from a large web corpus (Ljubešić and Klubička, 2014), encoded in a trie. Encoding the language model on the character level enables us to query it during the process of entering the current word. Encoding the language model in a trie assures a small memory footprint.

We calculate two evaluation metrics on the next word prediction task. The first metric is KSR – keystroke saving ratio, i.e. the ratio of keystrokes that did not have to be performed thanks to the prediction technology. The second metric is accuracy gain – the difference in word accuracy when the technology is used and when it is not. Namely, while using predictors of following words, errors can be omitted by selecting the intended word from the list of candidates before making a misspelling.

During this evaluation we take into account two variables. The first variable is the maximum number of candidates the next word prediction technology offers to the user. As the number of candidates increases, both evaluation metrics should increase as well. However, one should expect a negative impact of too many candidates being shown to the user, especially among participants with language disorders for whom reading is an issue. We expect for the positive impact of showing more candidates to fall off at some point.

The second variable we take into account is the clinical status of the subjects on whose text production we evaluate the technology. We differentiate between two levels: subjects with language disorders and healthy subjects.

The results are shown in Figure 1. Regarding the keystroke saving ratio, we can observe that it is to expect that healthy participants do make greater keystroke savings than participants with language disorders. Given that there is no significant difference in the number of misspellings on the first characters among these two groups, we assume that the observed difference is due to the fact that not all errors are corrected by using this technology, and having five times more spelling errors in the context on which we predict the following words surely has an impact. Furthermore, we can hypothesize that the text being produced by participants with language disorders has less of a natural flow and therefore the predictors of the following words, which use the already entered words for prediction, do not perform as good.

Regarding the optimal number of candidates to be shown to users, it should be set at around four candidates as at that point the ratio of keystroke savings does start to fall off. However, for optimizing this variable a set of experiments on live subjects should be performed.

Concerning the accuracy gain obtained through the use of this technology, among healthy subjects the absolute rise in accuracy as the number of candidates increases is much smaller than in speakers with language disorders. There is an accuracy gain of 0.3% if just one candidate is shown and 0.5% if ten candidates are shown. The error rate among those subjects is in general quite low, about 1%, so by using a predictor of the following word, the error reduction among healthy speakers on the word level is between 30% and 50%.

On the other hand, among subjects with language disorders, increasing the number of candidates shown does increase the accuracy gain quite visibly. While the accuracy gain is 1.2% with one candidate shown, by showing 10 candidates, it reaches 2.3%. Therefore overall error reduction among speakers with language disorders when using a predictor of the following word, in optimal conditions, is between 25% and 45%, just slightly lower than among healthy speakers.

Again, similar to the keystroke saving ratio, the accuracy gain starts to fall off around three to four candidates.

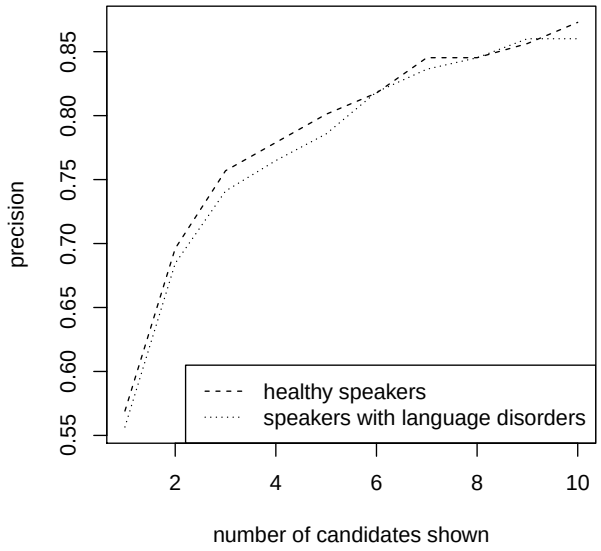

Figure 2: Evaluation results of the spelling corrector

6.2. Spelling Corrector

Our spelling corrector is based on an inflectional lexicon (Ljubešić et al., 2016), calculating the Damerau-Levenshtein distance of a potentially misspelled word to other words in the lexicon, and ranking the words satisfying the similarity criterion by using a language model trained on a very large web corpus (Ljubešić and Klubička, 2014).

We evaluate our spelling corrector by calculating the precision of the candidates shown on misspelled words. Precision is calculated on the level of a set of candidates where each set containing the right word is considered correct and the set not containing the correct word as incorrect.

We take into account the same two variables as in the evaluation of the predictor of the following word: the number of candidates shown to the user, and the two levels of language disorders status.

The results of this experiment are shown in Figure 2. Again, we can observe that the gain obtained by showing more candidates to the users starts to drop off at three candidates. At that point the precision of the candidate sets is around 75%, meaning that every fourth word does not have the correct form shown among the candidates.

Interestingly, in this experiment there is no difference to be observed among our two groups of subjects. The difference between the two groups observed in keystroke saving ratio and the lack of difference in the case of the precision in spelling suggestions can be explained by the fact that typing errors are five times more frequent among subjects with language disorders. These errors impact the predictor of following words negatively if they occur before the correct word is being suggested, but they do not impact the spelling corrector as the Damerau-Levenshtein distance distribution between the two groups is identical.

7. Conclusion

The corpus presented in this paper is a result of an 8 months long work of an interdisciplinary team consisting of speech and language pathologists, linguists, NLP scientists and information experts, and is, as such, a unique language resource, annotated on multiple levels, offering opportunities for different lines of research. However, its primary purpose was to provide data necessary for building a platform which would help speakers with various types of language disorders in producing written text and thus overcome one of the main obstacles in fulfilling their personal potentials in education and in the job market.

The analyses we presented here gave us important guidelines in designing such a platform. We have shown that, although making significantly more errors than healthy subjects, speakers with language disorders do neither make proportionally more errors of Damerau-Levenshtein distance higher than 2 nor make more errors on initial characters in the word, which confirms that traditional approaches of spelling correction and next word prediction will be applicable. Further on, we have searched for optimal number of candidates that are to be shown to the user using several evaluation metrics, keystroke saving ratio and accuracy gain for the next word predictor, and precision for the spelling corrector. In case of both presented technologies, the optimal number of candidates turns out to be 3 to 4. However, these numbers should be confirmed experimentally on live subjects.

Sad statistics reveal that not only the incidence of aphasia due to tumors or stroke gradually increases in the population, but the average age of people affected by this neurogenic disorder actually decreases. In this light, we find the presented results very promising and hope that the platform we are creating will help with the (re)integration of the population with language disorders into the society.

8. Acknowledgments

The research leading to these results has received funding from the European Regional Developmental Fund 2007-2013 under grant agreement No. RC.2.2.08-0050 (project RAPUT).

9. Bibliographical References

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Kuvač Kraljević, J. and Kologranić Belić, L. (2015). Grammatical and orthographic features in written language production of youth with specific language impairment. Poster presented at IV clinical linguistics international congress, Barcelona.

Kuvač Kraljević, J., Hržica, G., Olujić, M., Kologranić Belić, L., Palmović, M., and Matić, A. (2016). Sampling challenges of specialized spoken and written adult speakers corpora. In Kristina Cergol Kovačević et al., editors, *Proceedings of the 29th conference of Croatian Applied Linguistics Society: Applied Linguistc Research and Methodology*, pages 157–168, Zadar, Croatia. Srednja Europa, HDPL.

Kuvač Kraljević, J., Hržica, G., and Kologranić Belić, L. (in press). *Croatian Corpus of Non-professional Written Language*. University of Zagreb, Laboratory for Psycholinguistic Research, Zagreb.

Kuvač Kraljević, J., Matić, A., Kologranić Belić, L., and Olujić, M. (under revison). Written narratives of adolescents with specific language impairment: discourse analysis. *Journal of Communication Disorders*.

Ljubešić, N. and Klubička, F. (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.

Ljubešić, N., Klubička, F., Agić, Ž., and Jazbec, I.-P. (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 (LREC'16)*. European Language Resources Association (ELRA).

Miłkowski, M. (2007). Automated building of error corpora of polish. In Barbara Lewandowska-Tomaszczyk, editor, *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, volume 17 of *Łódź Studies in Language*, pages 631–639. Peter Lang.

Pedler, J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. PhD thesis, London University.

Rello, L., Baeza-Yates, R., Saggion, H., and Pedler, J. (2012). A first approach to the creation of a spanish corpus of dyslexic texts. In Nicoletta Calzolari, et al., editors, *LREC Workshop Natural Language Processing for Improving Textual Accessibility*, pages 22–27. European Language Resources Association (ELRA).

Rodrigues, P. and Rytting, C. A. (2012). Typing race games as a method to create spelling error corpora. In Nicoletta Calzolari, et al., editors, *LREC Workshop Natural Language Processing for Improving Textual Accessibility*, pages 3019–3024. European Language Resources Association (ELRA).

10. Language Resource References

Ljubešić, Nikola. (2013). *MULTEXT-East Croatian Morphosyntactic Specifications, revised Version 4*. 4.0-revised.

Mitton, Roger. (1985). *Birkbeck spelling error corpus*. University of Oxford Text Archive.

Wikipedia. (2015). *Wikipedia:Lists of common misspellings/For machines — Wikipedia, The Free Encyclopedia*.