Language Editing Dataset of Academic Texts

Vidas Daudaravičius
VTeX
Mokslininku st. 2a, Vilnius, Lithuania
vidas.daudaravicius@vtex.lt

Abstract

We describe the VTeX Language Editing Dataset of Academic Texts (LEDAT), a dataset of text extracts from scientific papers that were edited by professional native English language editors at VTeX. The goal of the LEDAT is to provide a large data resource for the development of language evaluation and grammar error correction systems for the scientific community. We describe the data collection and the compilation process of the LEDAT. The new dataset can be used in many NLP studies and applications where deeper knowledge of the academic language and language editing is required. The dataset can be used also as a knowledge base of English academic language to support many writers of scientific papers.

Keywords: Language editing, Academic texts, Dataset

1. Introduction

Language editing is the one of the last but not the least tasks in the publishing cycle of the scientific paper. The English is the main language to publish scientific papers. Therefore, writing of a scientific paper requires less effort for English native speakers and much more effort for researchers for whom English is the second language. The lack of tools for writing scientific papers in English is a formidable barrier to their being published, and to be seen in the scientific community, especially for the beginners. Nevertheless, the genre of scientific language is declarative and requires writing clearly, thoroughly and unambiguously. The language of a scientific paper should be fluent and free of grammar errors.

The use of determiners and prepositions is one of the toughest problems for non-native speakers, especially those living in a non-English speaking environment. The biggest obstacle for developing grammatical error correction systems has been the lack of availability of large, annotated corpora of texts that could be used as a standard resource for empirical approaches to grammatical error correction. The issues have been explored extensively in the literature (see Leacock et al. (2010)).

Recently, several shared tasks have been organised to work on grammar error correction (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013). They constitute a major step toward clarifying the possibilities of building novel grammar error correction technologies.

Several new datasets were published and made freely available recently. The first dataset is the CLC FCE Dataset (Yannakoudakis et al., 2011). The dataset contains 1,244 scripts produced by learners taking the First Certificate in English (FCE) exam, which assesses English at an upper-intermediate level. The scripts are either a letter, a report, an article, a composition or a short story, between 200 and 400 words, and are linked to meta-data about the question prompts, the candidate?s grades, native language and age. The overall size of the CLC FCE corpus is 423,850 tokens in length, and it is annotated with more than eight thousand grammar error corrections. The dataset was used in the HOO 2012 Shared Task on Preposition and Determiner Error Correction (Dale et al., 2012). Another new dataset was published in (Dahlmeier et al., 2013). The NUCLE corpus, the NUS Corpus of Learner English, is a collection of 1,414 essays written by students at the National University of Singapore (NUS) who are non-native speakers of English. The NUCLE corpus contains 1,220,257 tokens with 46,597 error annotations. The corpus was used in the CoNLL-2013 Shared Task on Grammatical Error Correction (Ng et al., 2013). Both datasets are comprised of texts written by learners of English taking exams. Therefore, the datasets contain particular types of errors typical for learners of English. The third dataset is the *Wikipedia Correction and Paraphrase Corpus*, WiCoPaCo, built by automatically mining French Wikipedia's revision history, (Max and Wisniewski, 2010). WiCoPaCo is a corpus of rewritings extracted from the revision history of Wikipedia. It includes spelling corrections, reformulations, and other local text transformations. The corpus focuses on local modifications made by human reviser and include various types of corrections and rewritings. As the aim of building WiCoPaCo was to extract local modifications, only rewriting of paragraph changing at most 7 words were taken into account. The corpus contains 146,595 spelling errors. The International Corpus of Learner English (ICLE) contains argumentative essays written by higher intermediate to advanced learners of English from several mother tongue backgrounds. The French Interlanguage Database (FRIDA) contains texts written by learners of French as a foreign language. It contains three separate sub-corpora: (a) texts written by English speakers; (b) texts written by Dutch speakers; (c) texts written by learners from various other mother tongue backgrounds. In addition to a raw text version, an error-tagged version of the corpus is also available. An extensive list of learners corpora with more than 100 entries is available on the internet[1].

The existing corpora with annotated corrections are either small, specialised or proprietary, and not available to the research community. These English learner corpora support the construction of tools for error correction in learners'

[1] http://www.uclouvain.be/en-cecl-lcworld.html

English. But this does not help for writing scientific papers in many domains, e.g. physics, mathematics, psychology, management, economics and other domains. A new dataset of Language Editing of Academic text is created to fill this gap, and which could be interesting for researchers in natural language processing. The dataset is based on texts processed at VTeX when providing science publishing services for the Springer and the Elsevier publishing companies in a broad range of domains: Physics, Mathematics, Economics and Management, Computer Science, Engineering, Statistics, Astrophysics, Chemistry and Human Sciences.

2. Language Editing at VTeX

VTeX provides LaTeX-based publishing solutions and data services to the scientific community and science publishers. Areas of services are: Typesetting, Copy editing, English language editing, Services for Authors, Electronic Publishing, Electronic Journal Management, Indexing, Full-text XML, Project Management. VTeX processes over 220 science journal titles and over 240 volumes of books comprising to more than 350 hundred pages each year. English language editing services are not requested by all journals, so many papers received no language revision.

Linguistic editing of scientific papers is a specific activity and there are no advanced tools for that. At the same time, the community of non-English speaking scientists is growing rapidly, so the percentage of papers written in poor-quality English is increasing. VTeX has unique archives of scientific texts before and after linguistic editing performed by linguistic editors. Modern methods of data mining and machine learning, well known in the area of natural language processing (NLP), have shown promising results. Which might be interesting not just to VTeX editors, but also to science publishers and the scientific community in general. The amount of language-copy-edited material, which passed the typesetting process at VTeX when providing publishing services for the Springer and the Elsevier, is large, and can be data-mined for development of new advanced tools for authors and language-editors. Therefore, VTeX is creating the new dataset, which is based on the production made in VTeX. 48 journal titles and books with 3,998 papers were selected. These papers were written by native and non-native English writers. Many papers have several co-authors with different English writing skills. Therefore it is difficult to handle information about the native language of writers. All selected papers were copy-edited by professional native English language editors with specialisation in the relevant domain. Each paper is edited by language editor only once. An experience and tradition of language editing vary from one editor to another, from domain to domain, and it is impossible to preserve the consistency of language edits. Therefore, it is expected to appear contradicting edits in some situation, and this should be taken into account when machine learning methods are applied to the data.

3. Document vs. paragraph

The data sources are scientific articles written by writers with different writing skills in English from the different countries, institutions and with the different professional experience. A common case is that many parts of the new papers are extracts from the old papers, and co-authors write new parts of the paper independently. Therefore, a paragraph is the most appropriate and consistent unit for the later analysis of language edits. Also, there is a problem to publish full articles and language editing data altogether. Full articles contain personal identity, which cannot be easily removed. Language editing data can reveal personal skills in English writing, which is usually confident personal information and should not be publicly available. The only way to reduce the risk of the personal identification is to take a paragraph as the largest text extract. In the light of these important concerns, we decided to use paragraph as the largest unit for the LEDAT.

4. The extraction of edits from the LaTeX documents

LaTeX is a common typesetting approach to high quality publishing, and it is widely used for writing scientific papers where complex and unified typesetting is important. The LaTeX documents include mathematical formulas, pictures, tables, listings, titles, citations, references, and other objects. All selected papers were encoded in LaTeX. A straight-forward way to extract texts from these documents is to use already available tools such as *detex* or *catdvi*[2]. *Detex* is a common approach to clean the LaTeX encoded

Domain	Source tokens	Target tokens	Edits	ParEdits	$\frac{\text{Edits}}{\text{ParEdits}}$	$\frac{\text{Source tokens}}{\text{ParEdits}}$
Physics	3,511,550	3,540,656	104,344	20,324	5.1	173
Mathematics	3,240,052	3,265,664	73,142	22,025	3.3	147
Economics/Management	1,278,951	1,288,421	29,054	8,655	3.4	148
Computer Science	851,802	856,899	16,798	5,696	2.9	150
Engineering	825,712	830,825	205,96	5,653	3.6	146
Statistics	553,764	558,642	14,213	3,963	3.6	140
Astrophysics	417,472	422,280	12,478	2,609	4.8	160
Chemistry	199,269	201,342	5,080	1,062	4.8	188
Human Sciences	17,931	17,893	313	149	2.1	120
Total	10,896,503	10,982,622	276,018	70,136	3.9	155

Table 1: Statistics of the Language Editing Dataset of Academic Texts.

[2] http://en.wikibooks.org/wiki/LaTeX/Export_To_Other_Formats

```
<parEdit id="421" domain="Physics">
<edit sourcePos="0" targetPos="0"><source>Further, the</source><target>The</target></edit>
<edit sourcePos="34" targetPos="25"><source>spectrums</source><target>spectra</target></edit>
<edit sourcePos="47" targetPos="36"><source></source><target>the </target></edit>
<edit sourcePos="57" targetPos="50"><source></source><target>the </target></edit>
<edit sourcePos="67" targetPos="64"><source></source><target>now </target></edit>
<edit sourcePos="166" targetPos="167"><source> years</source><target>-year</target></edit>
<edit sourcePos="188" targetPos="188"><source></source><target>the </target></edit>
<edit sourcePos="198" targetPos="202"><source></source><target>the </target></edit>
<edit sourcePos="296" targetPos="304"><source>the</source><target>a</target></edit>
<edit sourcePos="374" targetPos="380"><source></source><target>the </target></edit>
<edit sourcePos="384" targetPos="394"><source></source><target>the </target></edit>
<edit sourcePos="403" targetPos="417"><source>method of </source><target></target></edit>
<edit sourcePos="416" targetPos="420"><source></source><target> method</target></edit>
<edit sourcePos="455" targetPos="466"><source>of</source><target>for the</target></edit>
<edit sourcePos="468" targetPos="484"><source></source><target>the </target></edit>
<edit sourcePos="497" targetPos="517"><source>above</source><target>previous</target></edit>
<edit sourcePos="515" targetPos="538"><source></source><target>are </target></edit>
<edit sourcePos="575" targetPos="602"><source></source><target>the </target></edit>
<sourceText>Further, the global wavelet power spectrums of SPFNH and SPFSH are considered and shown
    as solid lines in Figure 5. Figure 5 shows that the spectral power of about 11 years period of
    both SPFNH and SPFSH is over the MATH confidence level line (dotted lines) (Torrence and Compo,
    1998). Moreover, the more precise value of the periodicity is obtained as 10.68 years for both
    SPFNH and SPFSH by using the method of FFT. Hence, the periods of about 11 years of SPFNH and
    SPFSH are identical with the above results and statistically significant. The periodicity of the
    SPFNH and SPFSH and the phase relationship between them are believable.
</sourceText>
<targetText>The global wavelet power spectra of the SPFNH and the SPFSH are now considered and shown
    as solid lines in Figure 5. Figure 5 shows that the spectral power of about 11-year period of
    both the SPFNH and the SPFSH is over the MATH confidence level line (dotted lines) (Torrence and
    Compo, 1998). Moreover, a more precise value of the periodicity is obtained as 10.68 years for
    both the SPFNH and the SPFSH by using the FFT method. Hence, the periods of about 11 years for
    the SPFNH and the SPFSH are identical with the previous results and are statistically significant
    . The periodicity of the SPFNH and the SPFSH and the phase relationship between them are
    believable.
</targetText>
</parEdit>
```

Table 2: An example of a parEdit.

text for the spell-checking purposes. The tool does not process LaTeX code, and it is difficult to predict the output of this tool as the output can contain nonlinguistic information. *Catdvi* outputs the text which is the result of the LaTeX compiler. The file should be processed with the LaTeX compiler before. The output keeps text formatting. Therefore, the text lines are justified, and it is difficult to identify paragraph boundaries. A text within tables and figures becomes as a regular text and not so acceptable for the later linguistics processing. So, we made a LaTeX to text converter to make the output more appropriate for the linguistic analysis. The main task is to keep linguistic formatting of the text, and to reduce the amount of nonlinguistic information, such as mathematical expressions, tables, and figures. The LaTeX code is preprocessed and macros are expanded. All mathematical expressions are replaced with MATH or MATHDISP. MATH stands for the in-line mathematical expressions, and MATHDISP stands for formulas typeset on a separate paragraph. Various citations, such as `\cite`, were replaced with CITE, and all references were replaced with REF. Tables and figures were replaced with TABLE and FIGURE, respectively. Each paragraph was placed on a separate text line (see Table 2) using LaTeX paragraph boundary detection style.

We applied the text extractor to all selected papers. Each paper has two versions: before language editing and after language editing. In this way, we processed 7996 documents. Next, we applied a diff algorithm to detect and align language edits on the single paragraph level. The aligned paragraph with language edits is referred to as a *parEdit*. We kept all paragraphs with the length between 100 and 5000 characters (including white-spaces).

We removed paragraphs that can easily identify papers or authors, like information on grants or addresses. We have manually checked and cleaned entries that contain one of the following strings: *support* and *grant* within the same paragraph, *acknowledg*, *keywords*, *grateful*, *pleasure*, *he is*, *she is*, *he received*, *she received*.

Finally, 80 percent of paragraphs were randomly selected and ordered, so that larger parts of text could not be easily recompiled. The selected data were used to compile the final dataset. The remaining 20 percent of the data is kept unpublished in case if new and unseen data is needed. The main statistics of the final dataset are presented in Table 1. More than 60 percent of the dataset includes texts from Physics and Mathematics, which are heavily loaded with mathematical expressions. The average paragraph length is 155 tokens but varies by domain. Long paragraphs tend to appear in Physics, Chemistry, Astrophysics. And short sentences tend to appear in Human Sciences and Statistics.

The diff algorithm was employed to detect language edits on the token level. All edit entries in the dataset were extracted automatically. Each edit includes the position of the edit source text and target text. The position is based on counting characters, not bytes. For instance, an XML character *&amp;* is counted as the length of one character. All edits are one of the two possible editing actions: insertion or deletion. Word or phrase order change is encoded with two independent edits: insertion and deletion. An example of a phrase order change and the annotation of edits is as follows:

Deletion	Insertion	Count	Deletion	Insertion	Count	Deletion	Insertion	Count
<empty>	,	55200	<empty>	of␣	550	towards	toward	323
,	<empty>	14975	a␣	<empty>	509	␣	,	322
<empty>	-	14798	’s	<empty>	499	.	:	315
␣	-	14320	<empty>	’	492	modelling	modeling	313
<empty>	the␣	10495	section	Sect.	491	that	which	301
-	␣	5243	Equation	Eq.	475	”	“	293
<empty>	␣	3995	,	.	470	;	,	292
␣	<empty>	3348	:	.	461	<empty>	it␣	285
<empty>	a␣	3329	<empty>	MATH␣	456	<empty>	–	281
Figure	Fig.	2938	which	that	455	equation	Equation	279
the␣	<empty>	2790	<empty>	is␣	441	<empty>	‘	277
<empty>	.	2782	(	<empty>	437	‘	”	274
:	<empty>	2535	’	<empty>	430	non-zero	nonzero	273
,	;	2178	equations	Eqs.	421	of␣	<empty>	265
-	<empty>	1844	<empty>	-,	417	<empty>	us␣	264
<empty>	:	1684	is	are	412	:	,	262
<empty>	)	1597	Fig.	Figure	398	figure	Figure	260
<empty>	and␣	1538	&	and	396	”	<empty>	254
Section	Sect.	1429	can not	cannot	390	<empty>	one␣	253
<empty>	(	1372	.	,	388	<empty>	to␣	248
<empty>	that␣	1304	␣	–	378	MATH	MATHth	236
equation	Eq.	1109	<empty>	”	378	the	The	235
.	<empty>	852	are	is	361	Figures	Figs.	235
behaviour	behavior	695	<empty>	;	356	it␣	<empty>	230
)	<empty>	682	figure	Fig.	353	that␣	<empty>	223
<empty>	an␣	677	<empty>	for␣	348	.	;	221
section	Section	636	,	:	343	<empty>	in␣	220
the	a	627	<empty>	by␣	338	Pomeron	pomeron	211
a	the	621	a	an	337	”	<empty>	210
MATH-th	MATHth	600	’	”	336	<empty>	be␣	203

Table 3: The top frequency list of language edits.

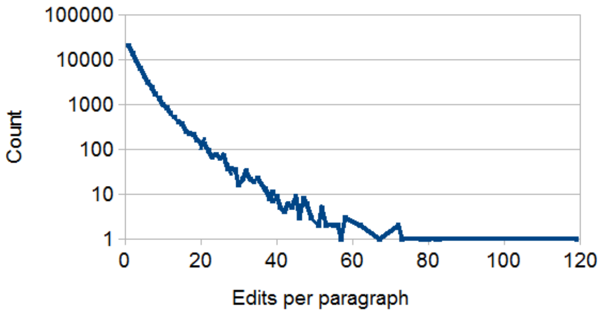

Figure 1: The distribution of the number of edits for each paragraph.

Insert: <edit><source></source><target>for the two models considered here and</target></edit>

Delete: <edit><source>for the two models considered here</source><target></target></edit>

Source text: The maximum likelihood estimates of all the marginal parameters, obtained when the competing risks are dependent, are given in Table REF for the two models considered here.

Target text: The maximum likelihood estimates of all the marginal parameters, for the two models considered here and obtained when the competing risks are dependent, are given in Table REF.

The average number of edits within a single paragraph is 3.9. Highly edited paragraphs are in Physics and Astrophysics. Less edited paragraphs are in Human Sciences and Computer Science (see Table 1). The top frequent edits in Table 3 show that the most frequent editing is comma insertion, which is about 20 percent of all edits. Surprisingly, insertion or deletion of a single space occurs on the top list also (3995 and 3348 times respectively). A common edit is the deletion of the space between text and a bracket, colon, semicolon, citation or a reference. Concatenation of several words into one compound or splitting a compound into several words requires insertion or deletion of a space. The distribution of the number of edits for each paragraph shows that some paragraphs are highly edited (see Fig. 1).

5. The types of language edits

The analysis of language edits of the dataset reveals three general types of language edits made by language editors:

Grammar corrections:

From: *On the other hand, as has been demonstrated, the sets MATH are in one-to-one correspondence to **group MATH elements**.*

To: *On the other hand, as has been demonstrated, the sets MATH are in one-to-one correspondence to **elements of the group MATH**.*

From: *with MATH and MATH rather well **describe** the invariant distribution of measured MATH's in p+p collisions␣.*

To: *with MATH and MATH rather well **describes** the invariant distribution of measured MATH's in p+p collisions.*

Lexicon and spelling:

From: *[...] the operators MATH and MATH are **he** left and right generators of MATH and [...]*

To: *[...] the operators MATH and MATH are **the** left and right generators of MATH and [...]*

From: *To **present** the spin operators of the group MATH, it is also useful to [...]*

To: *To **represent** the spin operators of the group MATH, it is also useful to [...]*

From: *The **irreps** of MATH are characterized by eigenvalues of two different [...]*

To: *The **irreducible representations** of MATH are characterized by eigenvalues of two different [...]*

Text cleaning:

From: *[...] with respect to the s.r.f.**.** That is, [...]*

To: *[...] with respect to the s.r.f. That is, [...]*

6. Data licencing

The LEDAT dataset is released under the Creative Commons BY NC SA licence, which allows to share alike, modify, adopt, remix the data for non-commercial purposes, and requires to give credits to this paper.

7. Conclusion

We have presented the LEDAT, which is significantly larger than other similar datasets. A wide variety and large number of language edits in the the LEDAT can be used in many NLP studies and applications where deeper knowledge of the academic language and language editing is required. The dataset can be used as a knowledge base of English academic language, and can support many writers in writing higher language quality papers. Several applications can be developed such as: automatic assessment of the language quality of a scientific paper; author self-education in English writing; professional computer aided tools for language editing.

The dataset represents a new initiative for academic communities, publishers and others, who are involved in the publication cycle. The initiative proposes to share data, which can increase our understanding of the language of scientific paper and how to write in good academic English. Larger amounts of similar data would yield new tools for automating language quality scoring, which is so eagerly awaited by many journal editors, conference organisers and authors.

8. Acknowledgements

We would like to acknowledge the Springer and the Elsevier publishing companies for their permission to use many scientific papers and to publish many text extracts. This work has been partially supported by the Lithuanian business support agency administrated European Union Structural Funds (project No. VP2-1.3-UM-02-K-04-019).

9. References

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Dale, R. and Kilgarriff, A. (2011). Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Dale, R., Anisimoff, I., and Narroway, G. (2012). HOO 2012: A report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada, June. Association for Computational Linguistics.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2010). *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool Publishers.

Max, A. and Wisniewski, G. (2010). Mining naturally-occurring corrections and paraphrases from Wikipedia's revision history. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., and Tetreault, J. (2013). The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August. Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.