Alignment-based Annotation of Proofreading Texts toward Professional Writing Assistance

Ngan L.T. Nguyen
University of Information Technology
Hochiminh, Vietnam
ngannlt@uit.edu.vn

Yusuke Miyao
National Institute of Informatics
Tokyo, Japan
yusuke@nii.ac.jp

Abstract

This work aims at constructing a corpus to satisfy such requirements to support research towards professional writing assistance. Our corpus is a collection of scientific work written by non-native speakers that has been proofread by native English experts. A new annotation scheme, which is based on word-alignments, is then proposed that is used to capture all types of inarticulations and their corrections including both spelling/grammatical error corrections and paraphrases made by proofreaders. The resulting corpus contains 3,485 pairs of original and revised sentences, of which, 2,516 pairs contain at least one articulation.

1 Introduction

Detection and correction of misspellings and grammatical errors have been recognized as key techniques for writing assistance, and have extensively been studied in natural language processing (NLP) (Whitelaw et al., 2009; Gamon, 2010; Tetreault et al., 2010; Park and Levy, 2011). However, correcting misspellings and grammatical errors, which can be performed by normal English native speakers, does not satisfy all the requirements of professional writing (Futagi, 2010). The core of the proofreading process, in reality, is paraphrasing inarticulations, which can only be done by expert proofreaders. Considering the two paraphrased sentences (1a) and (1b) below, we can see that sentence (1b) is likely to be considered better by most people (Williams and Colomb, 2010), although neither of them contains any misspellings or grammatical errors.

(1a) *The outsourcing of high-tech work to Asia by corporations means the loss of jobs for many middle-class American workers.*

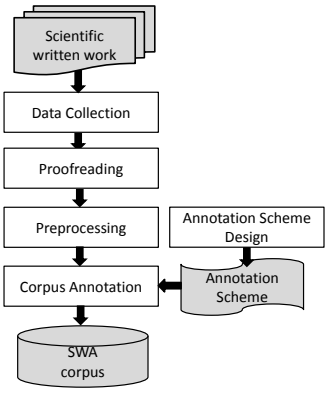

Figure 1: Methodology for corpus annotation

(1b) *Many middle-class American workers are losing their jobs, because corporations are outsourcing their high-tech work to Asia.*

(Williams and Colomb, 2010)

Most of the existing corpora are designed to capture errors in spelling and grammar, but they have not paid enough attention to paraphrasing.

We constructed a corpus that we called scientific writing assistance corpus (SWA), to support research on assistance with scientific-writing that captures all types of inarticulations, including those in both mispellings/grammar and paraphrasing. We have used the term *inarticulation* and *inarticulation correction* instead of *error* and *error correction* in this paper, to include in our task the paraphrasing, which is actually not errors.

Figure 1 overviews the methodology we proposed to construct the corpus. Scientific work written by non-native researchers or graduate students are collected (i.e., data collection, see Section 3), and this was then proofread by English native experts (i.e., proofreading). After that, we preprocessed the documents to convert them into a predefined format (i.e., preprocessing, see Section 3). Annotators with linguistic backgrounds were asked to strictly follow our annotation scheme, which had been designed to capture all types of

International Joint Conference on Natural Language Processing, pages 753–759,
Nagoya, Japan, 14-18 October 2013.

inarticulations (i.e., annotation scheme design, see Section 2).

Our corpus construction had several substantial advantages in comparison to the existing corpora such as the NUCLE (Dahlmeier and Ng, 2011), NICT_JLE (Izumi et al., 2004) and KJ corpora (Nagata et al., 2011). First, the proofreading process is separated from the annotation process. By doing this, both the writer and the proofreader were unaware of the construction of the corpus, so it could capture real articulations and corrections to these. Second, the alignment-based annotation scheme was employed in annotations to capture all types of articulation correction. This allowed us to annotate discontinuous paraphasing patterns, which were not neatly handled in other corpora. Third, paraphrases were captured, and were proved to be an important type of articulation correction for advanced learners.

The main contributions of this work are in the annotation of paraphrasing and its annotation method, in context of professional proofreading. Statistics for the SWA corpus was given in Section 4). We compared the grammatical errors annotated in the obtained corpus with those in the KJ corpus and NUCLE corpus, two popular corpora often used for research on grammatical error correction (Section 5), and performed an analysis of the paraphrases annotated (Section 6). Our analyses also show the potential of NLP research toward professional text revision.

2 Annotation scheme design

We extended the alignment-based paraphrase annotation scheme of Cohn et al. (2008) by categorizing the alignments into more fine-grained types (see Figure 2) to capture all types of inarticulation corrections. Figure 3 outlines example annotations to illustrate our annotation scheme. The alignments at the top level, are divided up into four broad types: Preserved, Metadata, Inarticulation Bi-alignment and Inarticulation Mono-alignment.

The Preserved type of alignments is the most trivial type that connects words with the same surface and function, e.g., *the*, *efficiency*, *various*, *methodologies* in Figure 3(A). Still, there are many cases where two words have the same surface form, but do not have the same functions in the original and the proofread sentences. For instance, the word *of* in the above example appears in both the sentences, but the two occurrences are

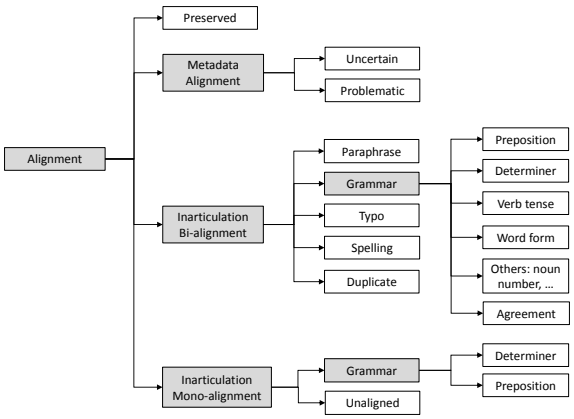

Figure 2: Proposed tagset. Categories in gray are used for classification but not for tagging.

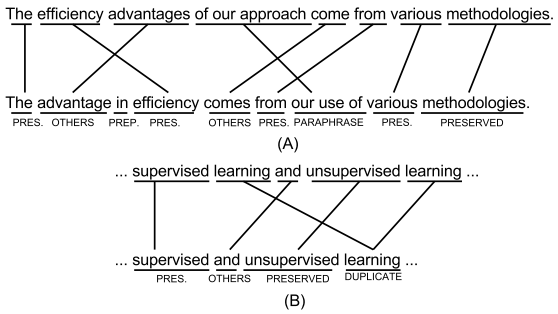

Figure 3: Annotations using our annotation scheme. Top has original texts, and bottom has the proofread text.

not aligned, because they modify different words, i.e., *approach* and *methodologies* in this case.

Inarticulation alignments including mono-alignments and bi-alignments are for capturing inarticulations and their corrections. The Grammar subtype of inarticulation alignments is not used for all types of grammatical errors as in the other annotation scheme, but is limited to some well-defined types of grammatical errors, which will be explained later in Section 2.1. The other subtypes are Duplicate, Spelling, Typo, and Unaligned, which will be explained in the following.

- **Duplicate**: A duplicate alignment connects words that appear once in the original sentence, but more than once in the proofread sentence, or vice versa. This tag captures the correction for articulations like the word *learning* in the example in Figure 3(B).
- **Spelling**: A spelling alignment is used for misspellings, e.g., *occured→occurred*[1]. This also includes the use of hyphens, e.g., *state of*

[1]The expression to the right of the arrow (→) is the preferred expression within context of writing

the art→state-of-the-art.

- **Typo**: The expression typo is a short form of typographical error, which refers to errors caused by typing mistakes. If annotators judge that the error is likely to be caused by a typing mistake, they should mark the errors as typo. Typo may be considered to be less important in writing assistance.
- **Unaligned**: An unaligned mono-alignment is used for words in the original sentence that have no correspondences in the proofread sentence, or vice versa.

Reordering of words are naturally captured by cross alignments, so we do not create a type for this. Punctuation marks are not annotated.

Besides, alignments have additional features to capture information that is specific to proofreading by humans. The current features are: Uncertain and Problematic. An alignment is marked as *uncertain* when the proofreader is not confident in the correction. This type is specific to the proofreading process. When the native proofreader is doubtful about his/her understanding of the original sentence, he/she will comment on it by stating *"I do not understand this,"* or *"This correction is a guess"*. An alignment is classified as *Problematic* when the annotators discover that the proofreader has made an erroneous correction. This happens when the proofreader misunderstands the author's intention. Although such situations can be rare, this tag is designed to offer a mechanism for annotators to provide feedback.

2.1 Grammar

Grammar-typed alignments connect a grammatical error in the original sentence with its correction in the proofread sentence. Grammatical errors in our annotation scheme are comprised of errors with determiners, prepositions, verb tenses, word forms, agreement, and others. They are tagged with the corresponding tags called Determiner, Preposition, Verb tense, Word form, Agreement, and Others. The Others type merges several specific subtypes of grammatical errors, including noun number, verb number, wh-word choice, or conjunction choice. Note that we do not use Others as a catch-all type. Except for Agreement, most of the subtypes of the Grammar type can be aligned well with the error types in the error taxonomies used by the existing corpora. The Agreement type is used to capture the number agreements of articles and nouns, genitives and nouns, or nouns and verbs, when a change in the number of one word forces us to change the number and form of another word.

2.2 Paraphrase

Any type of correspondence that cannot be classified into these types above is marked Paraphrase. In other words, Paraphrase is used as a catch-all type. Those errors that require complex corrections, i.e., corrections to phrase structures or sentence structures, which are not classified into the Grammar type, are captured with Paraphrase. We have followed the definition of paraphrases in the guidelines for paraphrase annotation by Callison-Burch et al. (2006): *"paraphrases convey the same meaning but are worded differently"*. We have two rules of thumb for the boundary of paraphrases: (1) shorter paraphrases are preferable (similar to (Callison-Burch et al., 2006)), and (2) a paraphrase alignment should not contain an alignment of other types in it.

3 Data collection and preprocessing

We collected scientific works that were written by seven authors with two language backgrounds Japanese and Vietnamese. The collected documents included different types of scientific publications such as short papers, full papers, and book chapters. We will use the terminology *document* to refer to a written work of any type. The collected documents belonged to two domains or fields of studies, which were computer vision (11 documents) and natural language processing (7 documents); and all were proofread by native English experts.

We then preprocessed these documents to convert them into a standard format. Non-text information such as figures and tables were removed. Format tags such as LaTeX's tags were also removed. We separated the original text and the proofread text for each document, and aligned the sentences in these two texts, so that a line in the original text corresponded to a line in the proofread text. We found that there were cases where a sentence in the original text should have been aligned with more than one sentence in the proofread text or vice versa. We allowed two or more sentences to be aligned in such cases.

4 Corpus annotation and results

We made use of Yawat, a web-based word-alignment annotation tool (Germann, 2008) to annotate the corpus. Yawat accepts text files containing pairs of aligned sentences as input. We applied a simple string-matching algorithm to produce default Preserved and Unaligned alignments for the corpus to save annotation time and effort.

The corpus was annotated by two annotators with linguistics background. The agreement between them was measured using the F1-score formula similarly to that by Cohn et al. (2008). *Atom-alignments*, or one-to-one alignments, were generated from the bi- and mono-alignments. An $M \times N$ multiple alignment would result in $M \times N$ atom alignments. We removed the preserved atom-alignments that were annotated by both annotators, because they occupied the majority of atom alignments but were not a meaningful indicator of inter-annotator agreement. Considering annotations by one annotator as gold annotations, we then calculated recall, precision, and F1-score over all the annotated alignments in the two versions of the SWA corpus. The overall F-scores with and without considering alignment types were 0.637 and 0.716 respectively. It can be seen that our inter-annotator agreement measure without considering alignment classification is comparable to those reported by Cohn et al. (2008) (0.71, 0.74, and 0.76, for the three datasets of MTC, Leagues, and News, respectively). This is reasonable because when alignment classification is taken into account, the annotation task is more difficult, so the inter-annotator agreement is lower.

A total of 4,686 Inarticulation alignments were annotated for 2,516 pairs of sentences in 18 documents. 69,738 (91.8%) of the total of 75,968 words in the corpus were annotated with Preserved alignments. Table 1 lists the ratios (%) of broad types of alignments. We can see that the Grammar errors, both in bi- and mono-alignments, occupy 58.1% of the total errors, which is not a surprise. Paraphrase alignments occupy a significant part, i.e., 29.3% of the total. These figures indicates that paraphrasing is an essential type for scientific writing; therefore, research on writing assistance should pay more attention to error correction by using paraphrasing.

The ratios of the subtypes of Grammar alignments are listed in the column named *SWA* (the name of our corpus) in Table 2. Out of all grammatical errors, determiners caused a lot of troubles for non-native writers from the Japanese and Vietnamese language backgrounds, even though the authors of the collected documents all had an advanced level of proficiency in English. This may be because of the difference between the characteristics of their background languages and the English language.

Alignment Type	Count	Ratio (%)
Paraphrase	1,372	29.3
Bi-Grammar	1,511	32.2
Typo	68	1.5
Spelling	308	6.6
Duplicate	13	0.3
Preserved	2	0.0
Mono-Grammar	1,212	25.9
Unaligned	200	4.3
TOTAL	4,686	100.0

Table 1: Statistics for all alignments (except for the Preserved type) annotated in the corpus

5 Cross-corpora comparison for grammatical errors

This section compares the grammatical-error annotations (Grammar alignments) in our corpus with those annotated in the KJ and NUCLE corpora. The Grammar types of errors in our scheme are restricted to well-defined types of grammatical errors. It would be interesting to analyze the differences in grammatical errors made by writers of the three corpora. The writers for our SWA corpus were graduate students and researchers in the field of computer vision and natural language processing, who could be considered to be advanced learners. The writers for the KJ and NUCLE were Japanese and Singaporean students, respectively.

As the three corpora used different annotation schemes, we created a mapping between compatible tags in the three tagsets to compare our corpus with theirs. This mapping is summarized in Table 3. The annotation scheme used for the KJ corpus, called KJ annotation scheme, was a simplified version of the NICT_JLE annotation scheme (Nagata et al., 2011). The definitions of types and marking schemes are basically similar in the two annotation schemes, but the KJ annotation scheme merges several subtypes into one type, for example, the

Type	KJ Count ($\times\alpha$)	KJ (%)	NUCLE Count ($\times\beta$)	NUCLE (%)	SWA Count	SWA (%)
Determiner	543 (726)	18.7	6,004 (641)	12.9	1,176	25.1
Preposition	377 (504)	13.0	7,312 (781)	15.7	547	11.7
Others	404 (540)	13.9	5,486 (543)	10.9	427	9.1
Verb tense	249 (333)	8.6	3,288 (351)	7.1	369	7.9
Word form	317 (423)	10.9	2,241 (239)	4.8	151	3.2
Agreement	146 (195)	5.0	1,578 (168)	3.4	53	1.1
TOTAL of Grammar	2,036 (2,723)	70.0	25,509 (2,723)	54.7	2,723	58.1
Total of all types	2,907	100.0	46,597	100.0	4,686	100.0

Table 2: Statistics for Grammar alignments in SWA in comparison with KJ corpus and NUCLE corpus with α = TOTAL$_{SWA}$/ TOTAL$_{KJ}$, β = TOTAL$_{SWA}$/ TOTAL$_{NUCLE}$

KJ	NUCLE	SWA
at	ArtOrDet	Determiner
prp	Wcip	Preposition
n_num, rel	Nn, Vform	Others
v_tns	Vt	Verb tense
aj, v_lxc	Wform	Word form
v_agr	SVA	Agreement

Table 3: Tagset mapping of KJ, NUCLE, and SWA for comparison. Note that only corresponding tags are mapped.

noun inflection, *noun case*, *noun countability* and *complement of noun* of the NICT_JLE annotation scheme, are merged into one type, the *noun lexical*. The KJ tagset contains 19 tags, fewer than the total number of 45 error tags in the NICT_JLE tagset (Izumi et al., 2004). The NUCLE tagset has more fine-grained tags than the KJ tag set (27 tags).

The four types Determiner, Preposition, Verb tense, and Agreement in our tagset have counterparts in the KJ tagset, which are *at* (article), *prp* (preposition), *v_tns* (verb tense), and *v_agr* (verb agreement) tags, and in the NUCLE tagset, which are *ArtOrDet* (article or determiner), *Wcip* (wrong collocation/idiom/preposition), *Vt* (verb tense), and *SVA* (subject-verb agreement). Note that subject-verb agreement is only part of the Agreement type in our annotation scheme (see Section 2). The counts for the Others type were sums of the *n_num* (noun number) and *rel* (relative) types for the KJ corpus, and of the *Nn* (noun number) and *Vform* (verb form) types for the NUCLE corpus. The Word-form figure of the KJ corpus was a sum of the *aj* (adjective), and *v_lxc* (verb lexical) types. As NUCLE has the exactly corresponding type called *Wform* (word form), so we used the count of this type in our comparison.

The comparison statistics are summarized in Table 2. We can see in this table that the ratios (%) of the totals of basic grammatical types over the totals of all annotated inarticulations, are significantly different for the three corpora, which correspond to 70.0%, 54.7%, and 58.1% for KJ, NUCLE, and SWA. The differences probably reflect the actual proficiency levels of the writers. Texts in KJ and NUCLE are written by college students, but they are not the same. This can be explained by the fact that NUCLE's college students are studying in Singapore, where English is used as an official language, while KJ's students are living in Japan, where English is not usually heard in daily life. The SWA's writers are also not living in an English-speaking environment, but they made fewer basic grammatical errors than KJ's students, which is reasonable because they have a higher proficiency level.

We normalized the count of each error type by using α and β listed in Table 2 to directly compare the three corpora in more detail. The normalized counts are in parentheses, next to the actual counts. To our surprise, the SWA's writers, who were scientific writers, make numerous determiner errors: 1,176 errors, compared to 726 (KJ) and 641 (NUCLE). KJ's students made fewer errors of this type than SWA's writers. This is possibly due to the difference in the complexity of the sentence structures used by the three groups of writers. KJ's students wrote very short sentences, while advanced learners tended to write those that were longer and more complex. Additional analyses of the sentence lengths and structures would clarify this further.

6 Analysis for paraphrase alignments

We carried out an analysis of the annotated Paraphrase alignments for understanding the challenges and possible solutions for research toward automatic proofreading (Table 4). For this anal-

Type	Examples of annotation	Count	%
1.Short-form ↔ Long-form	*PCA → principle component analysis*	2	0.6
2.Verb ↔ Prepositional phrase	*to collect → of collecting*	13	3.6
3.Relative clause ↔ Participle	*needed → that need*	5	1.4
4.Active ↔ Passive	*has not ... studied → has not ... been ... studied*	13	3.6
5.Anaphoric pronoun ↔ Referent	*this → the result*	22	6.1
6.Selection	*have → provide* *on the contrary → on the other hand*	131	36.4
7.Mis-use/ Addition	*frontal → the front of*	55	15.3
8.Unknown/ Simplification	*good point → advantage*	32	8.9
9.Complex	*It is better if ... are used → Using ... is better*	87	24.2
TOTAL		360	100.0

Table 4: Subtypes of Paraphrase alignments representing different confusing patterns of writers.

ysis, we randomly picked 20 Paraphrase alignments from each annotated document, and manually categorized them into nine subtypes that approximately represented different confusing patterns of writers.

The first five subtypes in the table are rather well-defined types. They were used for such transformations as between short- and long-form of acronyms, or relative clauses and their reduced forms, and so on. These well-defined paraphrases occupy 20.8 percent of all the total samples. The transformation between active and passive forms, and between anaphoric pronouns and their concrete forms could be challenging, because it requires correct interpretation of the event or entity being mentioned. For not-well-defined paraphrases, we classified them based on the number and the part-of-speech of the inclusive words.

The Selection subtype is for the replacement of a word with another word of the same part-of-speech, or an idiom with another idiom. There were several causes of this type of inarticulation. One cause was that the writers used less-formal or ambiguous words, which were inappropriate for scientific writing style. Another cause was that they selected a word which did not precisely describe the intended meaning, due to the interference by the writer's background-language or other reasons. The latter reason would be more challenging for automatic proofreading applications. Selection-typed paraphrasing is very important for writing assistance, not only because of its frequency but also it is a representative example of the increasing fluency of texts.

The Mis-use/Addition subtype is applied when a word in the original text is replaced with a sequence of several words in the proofread text. This often happens when the original words do not provide enough details, or mis-describe what the writers mean. Unknown/Simplification is the reverse subtype of Mis-use/Addition. This subtype indicated that non-native writers sometimes used long descriptions instead of compact words, such as *good point* instead of *advantage*. These two subtypes reveal the demand for techniques to simplify, or to provide more information to the text.

The Complex subtype is for many-to-many alignments. While the changes made by the other subtypes above are rather local, this subtype often required global changes at high levels of a sentence structure, such as those in the example *it is better if ... are used → using ... is better*. Previous studies on text revision have suggested that such changes are necessary for the coherence of a bigger discourse such as a paragraph or whole document (Williams and Colomb, 2010). How to make use of discourse information in automatic proofreading is an interesting issue of NLP studies using our corpus.

7 Conclusion

We described the SWA corpus, which was constructed to support studies on assistance techniques for professional writing. The traditional problem of error annotation was reformulated as a paraphrase annotation of pairs of the original and proofread sentences. This view inspired us to extend the alignment-based annotation scheme to be used for our annotation process. The comparison with two existing popular corpora revealed that grammatical errors made by different types of writers varied a great deal. The advanced writers tended to make more inarticulations that require paraphrasing.

The SWA corpus can be used as benchmark data for different tasks including grammatical error correction, paraphrase extraction, and automatic alignment, in context of proofreading. Further research should be carried out for paraphrasing techniques. The corpus is made available for research community on request basis.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *J. Artif. Int. Res.*, 38(1):135–187, May.

Chris Callison-Burch, Trevor Cohn, and Mirella Lapata. 2006. Annotation guidelines for paraphrase alignment, 12.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4):597–614, December.

Daniel Dahlmeier and Hwee Tou Ng. 2011. Grammatical error correction with alternating structure optimization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 915–923, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Dale and Adam Kilgarriff. 2010. Helping our own: The hoo 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, Dublin, Ireland.

Robert Dale and Adam Kilgarriff. 2011. Helping our own: Text massaging for computational linguistics as a new shared task. In *Proceedings of the International Natural Language Generation Conference 2011*, Nancy, France, September.

Yoko Futagi. 2010. The effects of learner errors on the development of a collocation detection tool. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, AND '10, pages 27–34, New York, NY, USA. ACM.

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing: a meta-classifier approach. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 163–171, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ulrich Germann. 2008. Yawat: yet another word alignment tool. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 20–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. Sst speech corpus of japanese learners' english and automatic detection of learners' errors. 28:31–48.

John Milton and Vivying S. Y. Cheng. 2010. A toolkit to assist l2 learners become independent writers. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1210–1219, Stroudsburg, PA, USA. Association for Computational Linguistics.

Y. Albert Park and Roger Levy. 2011. Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 934–944, Stroudsburg, PA, USA. Association for Computational Linguistics.

Benno Stein, Martin Potthast, and Martin Trenkmann. 2010. Retrieving customary web language to assist writers. In *Proceedings of the 32nd European conference on Advances in Information Retrieval*, ECIR'2010, pages 631–635, Berlin, Heidelberg. Springer-Verlag.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*.

Casey Whitelaw, Ben Hutchinson, Grace Y. Chung, and Gerard Ellis. 2009. Using the web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 890–899, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph M. Williams and Gregory G. Colomb. 2010. *Style: Lessons in clarity and grace*. Boston, MA: Longman.