Multilingual Wikipedia: Editors of Primary Language Contribute to More Complex Articles

Sungjoon Park[1], **Suin Kim**[1], **Scott A. Hale**[2], **Sooyoung Kim**[1], **Jeongmin Byun**[1], **Alice Oh**[1]
[1] School of Computing, KAIST, Republic of Korea
[2] Oxford Internet Institute, University of Oxford, Oxford UK
[1]{sungjoon.park, suin.kim, sooyoungkim, jmbyun}@kaist.ac.kr, [2]scott.hale@oii.ox.ac.uk, [1]alice.oh@kaist.edu

Abstract

For many people who speak more than one language, their language proficiency for each of the languages varies. We can conjecture that people who use one language (primary language) more than another would show higher language proficiency in that primary language. It is, however, difficult to observe and quantify that problem because natural language use is difficult to collect in large amounts. We identify Wikipedia as a great resource for studying multilingualism, and we conduct a quantitative analysis of the language complexity of primary and non-primary users of English, German, and Spanish. Our preliminary results indicate that there are indeed consistent differences of language complexity in the Wikipedia articles chosen by primary and non-primary users, as well as differences in the edits by the two groups of users.

Introduction

Many people around the world communicate in more than one language, both written and spoken. The exact nature of how multilinguals choose language in various contexts, as well as show varying degrees of proficiency in the multiple languages is not well understood. The reason is that it is difficult to observe natural uses of language at a scale large enough to quantify and study in depth. Wikipedia offers a great resource for studying multilingualsm, as there are many editors who edit multiple language editions (Hale 2014). This paper presents one of the first large-scale, quantitative studies of multilingualism using Wikipedia edit histories in multiple language editions.

Multilingualism online

We define multilingualism as the use of two or more languages, which is in line with the traditional definition of bilingualism from linguistics (Grosjean 2010). This definition does not mean that such a multilingual individual possesses native fluency in multiple languages; indeed, offline research shows that multilingual individuals rarely have equal and perfect fluency in their languages (Haugen 1969). We therefore expect a varying level of grammatical proficiency and complexity in the text contributed by multilingual users online. We refer to a user's most frequently edited language edition as the user's primary language edition, and all other editions that user edits are referred to as the user's non-primary language editions.

On many user-generated content platforms a large proportion of the content is generated by a small percentage of very active users (Priedhorsky et al. 2007; Kittur et al. 2007), and multilingual users overlap to some extent with this group of power users (Hale 2015). It may be that some users are so devoted to a platform or cause that they contribute content in multiple languages despite poor language proficiency.

In order to better understand the nature of content contributed by multilingual users on user-generated content platforms, we analyze edits by multilingual users on Wikipedia, the world's largest general reference work. Wikipedia is one of the top 10 websites in terms of traffic volume, and its articles are often among the top results for many search queries on Google. More fundamentally, Wikipedia content has impacts far afield of the encyclopedia itself as the content forms the basis for knowledge graph entries on Google and is used in algorithms ranging from semantic relatedness in computational linguistics (Milne and Witten 2008; Strube and Ponzetto 2006) to (cross-language) document similarity in information retrieval (Potthast, Stein, and Anderka 2008).

The first edition of Wikipedia launched in English in 2001, and was quickly followed by editions in other languages each operating independently. As the project matured, these editions have been integrated more closely with a global account system providing a single login across all Wikimedia sites and inter-language links connecting articles on the same concepts across languages. Nonetheless, there remain large differences in the content available in different languages with 74% of all concepts having an article in only one language (Hecht and Gergle 2010). Approximately 15% of active Wikipedia users are multilingual, editing multiple language editions of the encyclopedia (Hale 2014). These users are very active in their first (or primary) language, but make much smaller edits in their secondary (or non-primary) languages. Other than a small mixed-methods study of the contributions of Japanese–English bilingual users editing articles about Okinawa, Japan, (Hale 2015), little is known

about the content contributions of these users at larger scales or across different language pairs.

Therefore, we analyze the complexity of edits made by multilingual users in four aspects to explore the following research questions:

- Is it possible to quantify the language complexity of edits made by primary and non-primary language users?
- Do primary users tend to edit parts of the Wikipedia articles with higher language complexity than non-primary users?
- Do we observe more natural language in the articles after primary users' edits compared to the articles after non-primary users' edits?

In this paper, we suggest methods of quantifying language complexity in Wikipedia edits. We apply the proposed methods and present preliminary results.

Materials and Methods

In this section, we introduce the data collection process and operational definition of multilingual user in Wikipedia. Then, we discuss the three types of language complexity measures: basic features, lexical diversity, and syntactic structure.

Dataset

Metadata about edits to Wikipedia are broadcast in near real-time on Internet Relay Chat (IRC), and we begin with this edit metadata for the top 46 language editions of Wikipedia, including Simple English, from July 8 to August 9, 2013 as collected by Hale (2014). In contrast to the prior work by Hale, we retain all edits to articles even if the users marked the edits as "minor." The metadata includes article titles, language edition, timestamp, user ids, and urls to the content of each edit. The original data comprises 5,362,791 edits by 223,668 distinct users. We identify multilingual users from the metadata and retrieve the content of all edits by multilingual users from Wikipedia using the Wikipedia API.

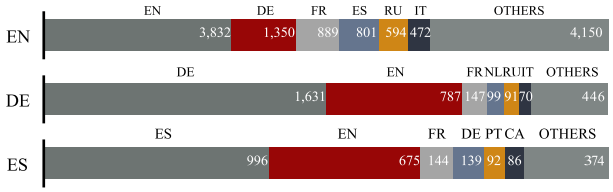

Figure 1: Proportion of editor's first language in each language edition. The English edition comprises the largest, yet the most various number of users by first language. 32.9% of the users who edited the English edition primarily edit English, compared to 49.9% of users in the German edition. Users having English as their primary language form the second-largest proportion of users in the Spanish and German editions.

	English	German	Spanish
# Editors	11,616	3,271	2,506
# Article edit sessions	237,849	120,123	69,557
# Edits	350,541	160,126	112,099

Table 1: Number of editors, article edit sessions, and edits for each language edition. On average, users had 20.5, 36.7, 27.8 article edit sessions and 30.2, 49.0, 44.7 distinct edits in English, German, and Spanish Wikipedia, respectively. The notable difference of the number of article edit sessions per user for each language edition implies there are different patterns of editing behavior between language edition.

Article Edit Session The most common way to measure edit activity in Wikipedia is by counting each edit that is created when a user clicks the "save page" button. However, counting edits does not accurately reflect authors' work because of individual differences in activity patterns (e.g., some users may commit a few large edits while other users may make a series of many smaller edits, saving the page more frequently as they work). There is a pattern of punctuated bursts in editors' activity, and we follow Geiger and Halfaker (2013) to use *edit session*, which measures the labor of Wikipedia editors. However, in this paper, we limit an edit session to one document, which we name *article edit session*. Also, we use *one hour* as cutoff between inter-session and between-session edit activities. After discovering inter-session edit activities, we aggregate all the distinct edits in an activity into an article edit session.

Finding Multilingual Users We first assume a user is able to read and write a language if the user participated in an article edit session in the language edition. After discovering article edit sessions, we define multilingual users as the users editing greater than or equal to two language editions. Using this definition, we identified 13,626 multilingual users with 1,113,004 article edit sessions, which comprises 1,595,556 distinct edits.

We can find that most multilingual users edit two or three language editions. 77.3% of multilingual users are bilingual, followed by 11.4% of trilingual and 4.1% of quadrilingual users. Users edited more than 10 languages account for 2.3% of all users, which we discard for this study because we regard them as either outliers or bots.

Further, we follow Hale (2014) to define a user's primary language as the most edited language with respect to the number of edit occurrences. Then, a user is *primary* in a language edition if the user's primary language equals to the language of the edition. Otherwise, the user is categorized as *non-primary* for the language edition. That is, a user is categorized as a primary language user in a language edition only once while regarded as a non-primary user as any other language editions.

We use three language editions of Wikipedia, English, German, and Spanish, for this study. Figure 1 illustrates the proportion of editor's first language in three language editions. English has 11,616, the largest number of unique users (see table 1). We found that users editing the En-

glish Wikipedia have the most various first languages. Only 32.9% of users edited English Wikipedia have their primary language in English, while 49.9% of German Wikipedia editors are German primary users. Also, English users takes the second-largest proportion in Spanish and German editions, implies English's role as a *lingua franca* in Wikipedia, as found in (Kim et al. 2014).

Data Processing For each edit, we retrieve the text of the article before the edit and after the edit and calculate the difference between these versions. For each article session, we extract the pair of changed paragraphs and convert the edit from Wiki markup to plain text. In this way, we retain only visible text from edits and discard all non-visible information including link and document structure. We regard an edit as minor if there is no visible change. We used NLTK sentence tokenizer and word tokenizer for Indo-European languages.

Controlling for Topics

There are articles with various topics in Wikipedia, and it is known that language complexity differs by the topics of its contents (Yasseri, Kornai, and Kertész 2012). Therefore it is necessary to cluster Wikipedia articles according to their topics in order to control them while comparing language complexity of editors using primary language and non-primary language.

Since it is hard to determine a single topic labels for a Wikipedia article with existing multiple categories, we cluster all of the articles included in the dataset using Latent Dirichlet Allocation (Blei et al. 2003). We set the number of topics to K=100 and employ online variational inference algorithm to estimate parameters with maximum iteration count 300. Using the 100-dimensional topic proportion vector for each document as feature vector, we cluster the articles again with DBSCAN algorithm for a single topic label. As a result, 20 topic clusters are discovered. To validate the cluster result, we calculate average distance to articles from cluster medoid over average distance to other clusters for each cluster computed as, in average, $0.59(\pm 0.22)$.

It is possible that the skewed user interest on different topic cluster would lead to inaccurate analysis. For instance, if a non-primary user tends to edit on low-complex topics or users having different first languages tend prefer the different topic, analysis might be affected due to the different interest on topics. To measure the user interest, we normalize the distribution of article edit sessions, to have the probability of edit for each cluster of sum to 1. We observe that the variety of interests within primary and non-primary language users, but overall indifference between groups. The complexity measures of primary and non-primary language groups were evaluated within the topics in advance, and then they were averaged to represent overall language complexity of each group to control inequivalent numbers of articles by topics.

Measuring Language Complexity of Edits

In an article edit session, each of the paragraphs before revisions (hereafter *before edits*) was paired with corresponding revised paragraphs (hereafter *after edits*). The following complexity measures are evaluated on each before and after edit paragraphs.

The computed language complexity measures for every edit pair (including before edits, after edits) were summarized with a single statistic for each editors to compare the complexity of edits produced by primary and non-primary groups. First, we chose *mean* value of complexity to comprehend about of actual edit patterns. Second, We also evaluated *maximum* value of complexity measures among edit pairs belongs to the same editor as a representation of the editor's linguistic ability to produce complex edits. This is mainly because all of the possible revisions in an article not always require editor's maximum linguistic ability. For this purpose, widely used central tendency measures (e.g. mean) would not reflect it properly. Summarizing edits with maximum is assuming that every editor showed their maximum linguistic ability to edit a paragraph at least once, which is reasonable.

Basic Features. First, we computed basic statistical complexity measures for *before edits* and *after edits* focusing on the length of edit paragraphs. The **number of characters, words, and sentences** are counted. Also, **number of unique words** is the number of word types appeared in the paragraph. **Average word length** is normalized number of characters by number of words and **average sentence length** is normalized number of words with respect to number of sentences in the edit paragraph.

Lexical Diversity. We additionally compute **entropy of word frequency** as a complexity measure which indicates uncertainty and surprise due to the newly appeared word in a paragraph. Yasseri et al. (2012) also interpreted entropy as measure of richness of word usage in a Wikipedia document.

We defined **word occurrence** and **word rank** to measure how an editor uses infrequent word in overall level, based on the entire dataset of every article edit session regarded as a repository of specific language. Word occurrence is frequency of n-grams, including unigram, bigram, and trigram, which is counted in the entire repository. The occurrence of every n-grams appeared within the edit paragraphs were averaged in a edit paragraph. Based on the occurrence, we also evaluated the word rank sorted in descending order by occurrence. That is, If a n-gram is frequent in the repository, it will have high rank. It can be intuitively expected that the word which is used more frequently is simpler in terms of language complexity as well as easier to use, so these measures are highly relevant to language complexity. The word rank of n-grams were averaged in a edit paragraph as with the word occurrence.

Syntactic Structure. Entropy of Parts Of Speech Frequency is computed for each edit paragraph pairs as well. Instead of analyzing the sequence of word itself, investigating sequence of Parts of Speech (POS) has some advantages. Not only it can ignore regarding extremely trivial edits (e.g., correcting misprints) as complex edits but also it can avoid the negative impact generated from bot-produced edits (Yasseri, Kornai, and Kertész 2012). In order to tag

the edits, we employ maximum entropy POS tagger (Ratnaparkhi 1996) trained on Penn Treebank corpus for English edits with Penn Treebank tagset (Marcus, Santorini, and Marcinkiewicz 1993). For German and Spanish edits, Stanford log-linear POS tagger (Toutanova et al. 2003) trained on NEGRA corpus with Stuttgart-Tübinger Tagset (STTS) (Skut et al. 1998) and AnCora corpus with its tagset (Taul et al. 2008) were used, respectively. Looking for diverse combinations of POS is good approach for detecting complex syntactic structure (Brian Roark 2007) since it measures the amount of information which indicates diverse usage of POS in the edits. We count the combinations of POS in each unigram, bigram, and trigram conditions and the normalized POS frequency distribution were used to compute information entropy for each edits.

Mean phrase length is the average number of words in each noun phrases (NP) or a verb phrases (VP) contained in every sentences in a edit. Since these phrases can emerge multiple times in a sentence and even embedded in a larger phrase, the phrase having the highest subtree in the entire tree was selected to compute the length only once for each sentence in order to evaluate them on overall sentence level. Length of NP is well-known measure of syntactic complexity because pre-modifiers and post-modifiers attached with it makes the phrase longer to hold and compress more information that increase complexity. The length of VP is measured to assess the complexity as well. **Mean parse tree depth** is the length of longest path in a parse tree. If two sentence length are the same, the larger tree depth could meaning that the sentence is more complex (Jieun Chae 2009). The depth of parse tree constructed from each sentences are averaged in an edit.

Prior to computing mean phrase length and parse tree depth, we used PCFG parser (Klein and Manning 2003) based on Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993) to construct parse tree for every sentences. Considering that parsing requires complete sentences, unlike POS tagging, for both before and after edits, we only include the edits containing at least 2 complete sentences into analysis while computing two measures. For this reason, they are evaluated only in case of English edits due to relative insufficiency of German and Spanish edits which are satisfying the constraints.

- **Basic Features**
  - Number of characters
  - Number of words
  - Number of unique words
  - Number of sentences
  - Average word length in characters
  - Average sentence length in words
- **Lexical Diversity**
  - Entropy of word frequency (unigram, bigram, trigram)
  - Average word rank (unigram, bigram, trigram)
  - Average word occurrence frequency (unigram, bigram, trigram)
  - Error rate of words (trigram)
- **Syntactic Structure**
  - Entropy of POS frequency (unigram, bigram, trigram)
  - Mean phrase length (Noun phrase, verb phrase)
  - Mean parse tree depth

Table 2: Language complexity measures for edit paragraphs. The measures in basic features are related to the overall length of edit paragraph. Other measures in lexical diversity focus on the usage patterns of words. The other measures relevant to syntactic structure examine the complexity of sentences based on the its POS tags and parse trees.

Results

In this section, we present the experimental results. We describe the evaluated complexity measures mainly focusing on the comparison between primary and non-primary user's language complexity in order to obtain implications of their editing behavior.

Basic Features. First, we present the actual edit patterns explored with basic features summarized with mean for each users. In English edits, primary users' basic features are higher than that of non-primary. The primary language user's single before edit comprises 139.2 characters equivalent to 33.6 words, 22.3 unique words, and 1.44 sentences. On the other hand, the average edit by a non-primary user is composed of 118.5 characters, which are 28.6 words, 19.3 unique words, and 1.31 sentences on average.

As with these result, after edits statistics are larger when in case of the primary language user's edit counted as 148.1 characters corresponding to 35.8 words, 24.0 unique words, and 1.6 sentences. Also, 132.8 characters, 32.2 words, 21.9 unique words, and 1.5 sentences composes single non-primary language user's after edit.

The tendency of primary language user's attempt to revise longer edits and preserve its length in after edits are also appearing in German and Spanish edits. Moreover, the results of basic features with maximum summarizing for users are shown in Fig. 2 that having even larger difference on every language editions between two groups.

These results indicate that there are some visible differences in part of the articles that users have modified and want to modify between two groups in English, German and Spanish editions. Through this, we can argue that there are significant difference between two groups in editor's ability, at least, to comprehend and attempt to revise longer edits.

Lexical Diversity. We found that the entropy of n-gram is always larger on the edits from primary language users, regardless of n and language editions. The top two plots in Figure 3 illustrate the entropy of primary and non-primary users on unigram, bigram, and trigrams for the English edition. The German edition shows the largest discrepancy of entropy between primary and non-primary users. These results imply primary language users are trying to revise more

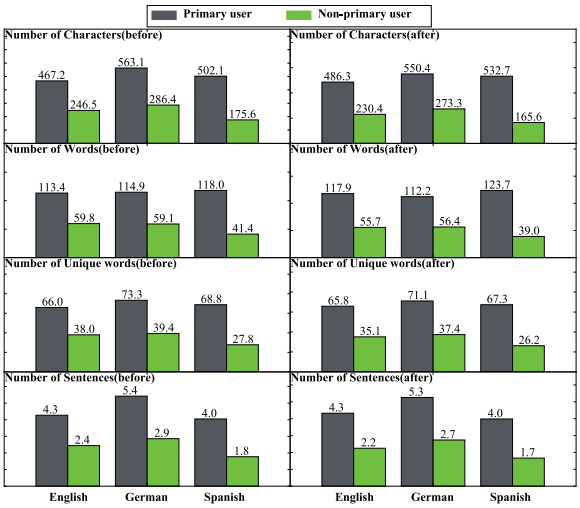

Figure 2: Four basic features (Number of characters, words, unique words, sentences) of before-edits (left) and after-edits (right) summarized with maximum for each users. As with Fig. 2 all of the measures of before-edits shows large difference between primary language users and non-primary which are reflecting the difference of ability relevant to read and comprehend complex edits. Also, results of after-edits indicate there are large difference remaining after edits.

diverse combinations of words and the diversity remains in corresponding after edits.

It is reasonable to interpret the higher entropy value in terms of higher diversity, but it is not meaning that they are trying to edit more *infrequently* used words. Interestingly, primary language users tend to revise the edit composed of more frequently used words with its variety combinations. This can be supported by the result of word occurrence and word rank measuring the usage of infrequent word sequence.

As a result, English primary users' average word occurrence of before-edits in English edition is 16,339,153, whereas it is 15,506,268 for non-primary editors. From this, it is natural that the average rank of word revised by primary users is 39,168, which is higher than that of English non-primary users, 44,987. The other language editions produce very similar tendencies as shown in Fig. 3. In addition, these tends remain in after-edits in each language edition.

Syntactic Structure. The entropy of POS sequence summarized with mean for each editors also differ by primary and non-primary users. In the English edition, English-primary users' before edits are evaluated as 1.146/bit, 1.542/bit, 1.626/bit while English non-primary's edits were 1.073/bit, 1.415/bit, 1.463/bit by unigram, bigram, and trigram, respectively. When moving on to estimate editor's ability with maximum case, the between group difference is dramatically increasing that the values are 2.057/bit, 2.957/bit, 3.284/bit in each n-gram condition of English-primary, whereas 1.571/bit, 2.160/bit, 2.317/bit in English non-primary on English articles. These tendencies are also shown in other language editions.

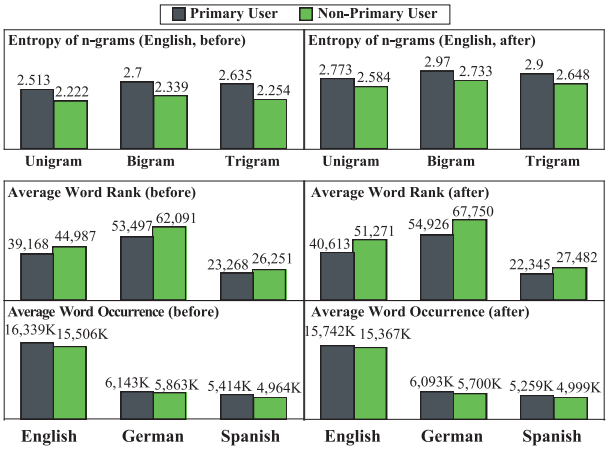

Figure 3: Error rate of primary and non-primary users. Compared to non-primary users, primary users tend to edit paragraphs with lower error rates (left). In all three language editions, the discrepancy between primary and non-primary users increases after their edits (right). On average, primary users' edits decrease the error rate while non-primary users tend to use more infrequent trigrams.

As with previous results, the primary language users' entropy values are slightly higher than non-primary users when it is averaged for each editors, and the difference increases when it is maximized for editors. It is confirmed again with complexity measures related to parse trees that shown below, Table 3.

	Before-edits		After-edits	
	P	NP	P	NP
Avg:NP length	11.87	11.81	11.93	11.94
Avg:VP length	15.40	15.03	15.40	15.13
Avg:Parse Tree depth	10.89	10.72	10.89	10.73
Max:NP length	18.19	13.86	18.23	14.02
Max:VP length	22.92	17.50	22.89	17.64
Max:Parse Tree depth	13.28	11.53	13.27	11.53

Table 3: Complexity measures derived from parse trees in before and after-edits of English edition. Each measures of P (primary) and NP (non-primary) language users are summarized with Avg (mean) or Max (maximum) for each editors.

Discussion and Future Work

Our preliminary findings suggest that multilinguals in Wikipedia show relatively high levels of proficiency in their primary languages. We find repeatedly that the majority of non-primary language edits are relatively short and simple, though many of them are just as long and complex as the primary language edits. We plan to conduct more analysis for future studies, and we project that these results will serve as an insightful starting point for large-scale quantitative research on naturally-occurring use of multiple languages.

Our results have important implications regarding the extent to which multilingual users may transfer information

between different language editions of the encyclopedia. While there are no doubt some large and complex edits by users in their non-primary languages that required genuinely high levels of multilingual proficiency, we find that the majority of non-primary language edits are small and simple in terms of language complexity. We further find that users are more likely to edit grammatically simpler parts of articles in their non-primary languages indicating that language complexity may form a barrier.

Also, these findings suggest that many of the users editing multiple language editions of Wikipedia may have a relatively low levels of proficiency in their non-primary languages. A good proportion of these users may be so-called power users who are very active on the platform (Priedhorsky et al. 2007; Kittur et al. 2007). These users may be making edits in multiple languages more out of dedication to Wikipedia and its cause than true multilingual proficiency (Hale 2015). Given that only roughly 15% of all Wikipedia editors edit multiple editions of Wikipedia (Hale 2014), our findings of low levels of non-primary language proficiency among a large proportion of these editors indicate there is significant work to be done in recruiting and encouraging multilingual contribution among truly proficient bilingual editors on Wikipedia if these users are to play any major role in the transfer of information between languages.

Meanwhile, The contribution to articles could be investigated by analyzing the difference between text before and after edit which are derived directly from the revised string. Also, *delta* defined in terms of difference of complexity measure between them could be explored independently. Not only these approach to Wikipedia but also various research questions could be investigated and we hope to tackle to the questions in the future works.

Acknowledgements.

We acknowledge support from MURI grant #504026, ARO #504033, NSF #1125095, and ICT R&D program of MSIP/IITP [10041313, UX-oriented Mobile SW Platform].

References

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and Lafferty, J. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:2003.

Brian Roark, Margaret Mitchell, K. H. 2007. Syntactic complexity measures for detecting mild cognitive impairment. *BioNLP '07 Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing* 1–8.

Geiger, R. S., and Halfaker, A. 2013. Using edit sessions to measure participation in Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 861–870. ACM.

Grosjean, F. 2010. *Bilingual: Life and Reality*. Harvard University Press.

Hale, S. A. 2014. Multilinguals and Wikipedia editing. In *Proceedings of the 6th Annual ACM Web Science Conference*, WebSci'14. New York, NY, USA: ACM.

Hale, S. 2015. Cross-language Wikipedia editing of Okinawa, Japan. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI 2015.

Haugen, E. 1969. *The Norwegian Language in America: A Study in Bilingual Behavior*. Indiana University Press.

Hecht, B., and Gergle, D. 2010. The Tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, CHI '10, 291–300. New York, NY, USA: ACM.

Jieun Chae, A. N. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. *Proceedings of the 12th Conference of the European Chapter of the ACL* 139–147.

Kim, S.; Weber, I.; Wei, L.; and Oh, A. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 243–248. ACM.

Kittur, A.; Chi, E.; Pendleton, B. A.; Suh, B.; and Mytkowicz, T. 2007. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Presented at alt.CHI, ACM SIGCHI Conference, San Jose, CA, USA.

Klein, D., and Manning, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of theE Association for Computational Linguistics*, 423–430.

Marcus, M. P.; Santorini, B.; and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics* 19(2):313–330.

Milne, D., and Witten, I. H. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*, WIKIAI 2008. Chicago, IL, USA: AAAI.

Potthast, M.; Stein, B.; and Anderka, M. 2008. A Wikipedia-based multilingual retrieval model. In Macdonald, C.; Ounis, I.; Plachouras, V.; Ruthven, I.; and White, R., eds., *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 522–530.

Priedhorsky, R.; Chen, J.; Lam, S. T. K.; Panciera, K.; Terveen, L.; and Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work*, GROUP '07, 259–268. New York, NY, USA: ACM.

Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speech tagging.

Skut, W.; Brants, T.; Krenn, B.; and Uszkoreit, H. 1998. A linguistically interpreted corpus of german newspaper text. In *In Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, 705–711.

Strube, M., and Ponzetto, S. P. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 2 of *AAAI'06*, 1419–1424. AAAI Press.

Taul, M.; Mart, M. A.; Recasens, M.; and Computaci, C. D. L. I. 2008. Ancora: Multi level annotated corpora for catalan

and. In *Spanish. 6th International Conference on Language Resources and Evaluation, Marrakesh*.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, 173–180. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yasseri, T.; Kornai, A.; and Kertész, J. 2012. A practical approach to language complexity: A Wikipedia case study. *PLoS ONE* 7(11):e48386.