AN ANALYSIS OF GRAMMATICAL ERRORS IN NON-NATIVE SPEECH IN ENGLISH

John Lee, Stephanie Seneff

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139 USA
{jsylee,seneff}@csail.mit.edu

ABSTRACT

While a wide variety of grammatical mistakes may be observed in the speech of non-native speakers, the types and frequencies of these mistakes are not random. Certain parts of speech, for example, have been shown to be especially problematic for Japanese learners of English [1]. Modelling these errors can potentially enhance the performance of computer-assisted language learning systems.

This paper presents an automatic method to estimate an error model from a non-native English corpus, focusing on articles and prepositions. A fine-grained analysis is achieved by conditioning the errors on appropriate words in the context.

Index Terms— computer-assisted language learning, second-language acquisition, grammar checking

1. INTRODUCTION

It is widely recognized that the usage of certain function words in English is difficult to master. For example, the misuse of articles and prepositions are among the most frequent types of grammatical errors in the Japanese Learners of English (JLE) corpus [1]. To provide aid to students of foreign languages, there has recently been much research on building computer-assisted language learning (CALL) [2] as well as grammar checking [3] systems.

The grammatical errors made by non-native speakers may be influenced by their first language, and are hardly random. Understanding the nature of these errors, and the contexts in which they occur, can be helpful to grammar checking systems, just as pronunciation tutoring systems have benefited from adaptation of their acoustic models with non-native speech corpus (e.g., [4]). However, to our knowledge, there has not been any reported effort to build non-native language models automatically from corpora.

This paper illustrates a step towards this direction of research, by first motivating the utility of such a model and reviewing related work (§2), then describing the data (§3), and finally presenting error analyses on articles and prepositions (§4-5). The notation [$\langle crr\rangle \rightarrow \langle err\rangle$] will be used to indicate that the correct word, $\langle crr\rangle$, is mistakenly replaced with the word $\langle err\rangle$ by the non-native speaker.

2. MOTIVATION & RELATED WORK

A number of NLP applications can benefit from a detailed analysis of grammatical errors made by non-native speakers.

2.1. Parsing Non-native Sentences

Grammar checkers and CALL systems must be able to process ungrammatical sentences, and provide high-quality feedback. Various strategies have been adopted, including the recognition of specific errors with hand-crafted templates [5], embedding error production rules in context-free grammars [6], and paraphrasing via a semantic representation [7]. In these approaches, the grammatical errors to be covered must be anticipated in advance; typically, the set of errors are compiled using anecdotal observations from domain experts. A systematic, corpus-based analysis can help harvest errors that might otherwise be overlooked.

In some settings, it might be desirable to selectively present feedback to the user. An error model, possibly personalized and updated on-line [8], can help prioritize and tailor the feedback.

2.2. Estimating Confidence in NLG

To eliminate the need to anticipate all errors, a class of grammar checkers adopts an approach centered on natural language generation (NLG) techniques. Using NLG models trained on a large body of correct English text, they predict the correct word to be used given a context, e.g., whether an article should be placed in front of a noun [9], and, if so, which one. This approach has also been applied to prepositions [10].

The broader coverage achieved by this approach often comes with the price of a lower precision. One way to improve precision is to compute a confidence measure, proposing a correction only when the measure exceeds a threshold. The measure can be, for example, the difference between the scores of the top- and second-place candidates given by the NLG component [10].

This kind of measure may potentially be sharpened by exploiting known tendencies of the author of the text. Suppose the input sentence is "*We had *the dinner at a restaurant*", and the NLG system proposes both "*null*" and "*a*". If an error model informs us that, in the context of the noun "*dinner*", the author is more likely to commit the error [*null*→*the*] than [*a*→*the*], then the deletion can be made with more confidence.

2.3. Simulating Non-native Sentences

A recent approach in grammar checking is to adapt statistical machine translation techniques, which generally rely on large parallel corpora of the source (i.e., non-native texts) and target (i.e., corrected version of those texts) languages. Since these resources are not yet available in large quantity, they are simulated by introducing errors into well-formed text. For instance, errors are inserted into mass nouns [11] based on patterns observed in a non-native corpus. Since the "translation" model is induced by the simulated data, the simulation quality is crucial. The quality depends on both the frequency and authenticity of the errors.

Without *a priori* knowledge of the error frequency, an equal number of grammatical and ungrammatical sentences were repre-

Type	Percent	Error	Percent	Example
Del	69.4%	[*a*→*null*]	41.1%	Three people go to
		[*the*→*null*]	28.4%	see [*a*→*null*] **movie**
Ins	17.3%	[*null*→*the*]	12.9%	We had [*null*→*a*]
		[*null*→*a*]	4.4%	**dinner** at a restaurant
Sub	13.3%	[*a*→*the*]	9.8%	How about going to
		[*the*→*a*]	3.5%	see [*a*→*the*] **movie**

Table 1. *Relative frequencies of deletions (Del), insertions (Ins) and substitutions (Sub) of articles, out of a total of 3382. Each error type is broken down into the specific errors. An example is drawn from the noun most frequently involved in each type.*

sented in the "source language" in [11]. A study on restoring missing articles [12] created a series of training datasets, in which progressively more articles were removed. These datasets yielded models with varying degrees of aggressiveness to suit input texts of different qualities. In both of these cases, an error model could have helped optimize the error frequency in the simulated training data.

An error model may also improve the authenticity of the simulated data. For example, rather than randomly removing articles [12], the model would bias towards nouns whose articles are most likely to be omitted by non-native speakers. Furthermore, authentically simulated data can be easily turned into cloze items [13], for the student's own practice and assessment.

3. DATA

While many detailed analyses have been performed on a small number of subjects (e.g., [14]) or a specific grammatical construction (e.g., [15]), few large-scale non-native corpora are available. Most, including the International Corpus of Learner English (ICLE) [16] and the Chinese Learner English Corpus [17], have very limited annotation. One exception is the Japanese Learners of English corpus (JLE) [1], on which our analysis is based.

The JLE corpus consists of transcribed interviews at the Standard Speaking Test (SST), an English-language proficiency test conducted in Japan. For 167 of these transcripts, grammatical errors have been annotated. We segmented [18] the turns of the interviewees in these transcripts, yielding 15637 sentences, with over 153K words. Our analysis in the rest of the paper is based on parse trees automatically produced by [19].

Some examples of article and preposition errors drawn from this corpus are shown in Tables 1, 3 and 5. To the extent that these errors are specific to native Japanese speakers, the utility of the acquired error model may be limited outside this linguistic community. However, our analysis techniques can potentially be applied to corpora from other communities as they become available.

4. ANALYSIS ON ARTICLES

A noun may be preceded by a determiner, most commonly an article, i.e., "*a*", "*an*"[1] or "*the*". In the JLE corpus, nouns have no determiner ("*null*") 41% of the time, and have "*the*" and "*a*" 26% and 24% of the time, respectively. The remaining 9% are other determiners, the majority of which are quantifiers such as "*some*", "*that*", "*this*", and "*those*".

[1] The distinction between "*a*" and "*an*" can be easily resolved and is not considered further. Both will henceforth be represented as "*a*".

Error Likelihood	
Context	**Example**
⟨*a*,**theater**⟩	Let us go to [*a*→*null*] movie **theater**
⟨*a*,**area**⟩	It's [*a*→*null*] residential **area**
⟨*a*,**concert**⟩	They offered [*a*→*null*] free **concert**
⟨*a*,**club**⟩	I want to belong to [*a*→*null*] basketball **club**
⟨*the*,**guitar**⟩	I like to play [*the*→*null*] **guitar**

Error Likelihood weighted w/ Frequency in General Domain	
Context	**Example**
⟨*the*,**market**⟩	... is the number one brand in [*the*→*null*] Japanese shampoo and conditioner **market**
⟨*the*,**team**⟩	[*The*→*null*] England football **team** has very famous and good players
⟨*a*,**company**⟩	She works for [*a*→*null*] real estate **company**
⟨*a*,**day**⟩	it was [*a*→*null*] rainy **day**
⟨*null*,**people**⟩	I think [*null*→*the*] young Japanese **people** think it's cool
⟨*null*,**one**⟩	I'm going to visit [*null*→*a*] **one** of my friends

Table 2. *The top table lists article-noun pairs with the highest error likelihoods (§4.1.1). The bottom table lists, for each article, the top two pairs when the likelihood is weighted with frequency in a general domain (§4.1.2).*

4.1. Confusion among Articles

The overall distribution of deletions, insertions and substitutions of articles, listed in Table 1, shows that deletion is the overwhelming type of error. This may be expected, since there is no functional equivalent of articles in the Japanese language. Among deletion errors, "*a*" is more often deleted, even though it appears almost as frequently as "*the*".

It would not be reasonable, however, to assume that non-native speakers omit articles randomly, regardless of context. When conditioned on the head noun of the article, the error type is no longer always dominated by deletions. For example, with the word "*police*", deletions are less prevalent than substitutions, e.g., "*I called [the→a] police*".

The noun that is most frequently associated with each error type, based on absolute counts, is shown in Table 1. This "top offender" list is inevitably influenced by the particular vocabulary distribution of the corpus. The fact that the words "*dinner*" and "*movie*" emerge on top likely has to do with the conversation topics at the examination. This issue may be addressed in two ways, one geared towards grammar checkers (§4.1.1), the other towards CALL systems (§4.1.2).

4.1.1. Error Likelihood

To decide whether to propose a correction for the article, it would be useful to measure how error-prone the head noun is. One could start by normalizing the article-error count of the noun by the number of times the noun appears in the corpus. However, its error-proneness may vary significantly depending on the article expected. For example, the noun "*place*" has the correct article 62% of the time overall, but for the subset with the article "*a*", the figure drops to 26%.

Thus, the noun by itself would not suffice as the *context* for article errors; it should rather be considered in conjunction with its expected article. The error likelihood of an article-noun pair is, then,

Type	Percent	Error	Percent	Example
Del	53.8%	[*in*→*null*]	14.6%	A car is parked
		[*at*→*null*]	8.3%	[*on*→*null*] the
		[*on*→*null*]	7.5%	**side** of the road
Sub	46.2%	[*on*→*in*]	5.1%	I study [*at*→*in*]
		[*at*→*in*]	3.7%	the **university**
		[*in*→*at*]	2.8%	

Table 3. *Relative frequencies of deletions (Del) and substitutions (Sub) of prepositions in adjunct prepositional phrases, out of a total of 780. An example is drawn from the complement most frequently involved in each error type.*

Error Likelihood	
Context	Example
⟨*along*,**street**⟩	I walk [*along*→*null*] the **street**
⟨*in*,**team**⟩	One of the most famous baseball players in Japan [*in*→*at*] the same **team**
⟨*on*,**birthday**⟩	I will go to your place [*on*→*at*] my next **birthday**

Error Likelihood weighted w/ Frequency in General Domain	
Context	Example
⟨*at*,**university**⟩	I studied [*at*→*in*] the **university**
⟨*at*,**end**⟩	[*At*→*In*] the **end** of the month, ...
⟨*for*,**year**⟩	I have taken lessons [*for*→*null*] about ten **years**

Table 4. *The top table lists preposition-complement pairs with the highest error likelihoods (§4.1.1). The bottom table lists the top candidates after weighting with frequency in a general domain (§4.1.2).*

the number of times the pair contains an article error, divided by its total number of appearances.

The article-noun pairs with the highest error likelihoods are shown in Table 2. Reflecting the overall tendency of underusing articles, the top pairs tend to have the article "*a*" or "*the*", while the ones with "*null*" dominate the bottom of the list (not shown). The most error-prone pair with "*null*" is ⟨*null*,**news**⟩, e.g., "*I read [null→a] news on the Internet*". In grammar checkers, for predicting the likelihood of a particular error, these likelihoods can be easily turned into $P(\langle err\rangle|\langle context\rangle)$ by specifying the identity of the error given each context.

4.1.2. Frequency in General Domain

Of most interest to users of CALL systems are those article-noun pairs that are not only error-prone, but also common in everyday usage. For the latter criterion, the AQUAINT newswire corpus is utilized to estimate the "unigram" probability of each pair, i.e., the proportion of that pair among all article-noun pairs in the corpus. When multiplied with the error likelihood (§4.1.1), the product may be interpreted as the probability of that pair occurring with an article error, had the AQUAINT corpus been written by a non-native speaker.

The top two pairs with the highest estimated frequencies for each article are listed in Table 2. These words tend to be both common and susceptible to article errors.

4.2. Confusion between Articles and non-Articles

Past research has focused exclusively on confusions among the articles. Although confusions between articles and other determiners are less numerous, they also exhibit some interesting trends. The single most frequent error is the use of "*some*" in place of an indefinite article, e.g. "*She selected [a→some] tie for him*". Not far behind are errors with possessive pronouns, e.g. "*I believe you understand [the→my] reason why I can't join your party*", and in the reverse direction, "*I like [my→the] children*". Other frequent errors are [*the*→*that*], and [*the*→*this*].

5. ANALYSIS ON PREPOSITIONS

The most frequent prepositions in the JLE corpus are "*in*" (23%), "*of*" (18%), "*for*" (12%), "*on*" (8%), "*with*" (8%) and "*at*" (8%). A preposition "expresses a relation between two entities, one being that represented by the prepositional complement, the other by another part of the sentence." [20] The *prepositional complement* is typically a noun phrase under the prepositional phrase (PP). The "another part of the sentence", also known as the *lexical head*, can be a preceding verb, noun or adjectival phrase. In Table 3, the bolded words in the examples are prepositional complements, while in Table 5, they are lexical heads.

To determine which one of these two entities provides a better context for preposition errors, a useful criterion is the argument/adjunct distinction[2] for PPs. Generally speaking, a preposition in an argument PP is closely related to the lexical head, serving as its argument marker. For example, in the sentence "*She looked at a monkey*", "*at a monkey*" is an argument for its lexical head, "*look*". A preposition in an adjunct PP is a modifier of the lexical head, and is less closely related to it. For example, "*at night*" is an adjunct modifier for the lexical head "*came*" in "*She came at night*". In principle, "arguments depend on their lexical heads because they form an integral part of the phrase. Adjuncts do not." [21].

These distinctions are not always clear-cut but, for arguments, the lexical head generally gives a better context for the preposition. Given "*She looked [at→null] a monkey*", the error seems specific to "*look*", and could have occurred with any other animal. In contrast, for adjuncts, the appropriate context word is the prepositional complement. Consider the error in the sentence "*She came [at→null] night*". It appears reasonable to assume that it is tied to "*night*", and could have occurred with verbs other than "*come*".

5.1. Adjuncts

In Japanese, case particles can play a role similar to English prepositions, although no simple mapping is possible [22][3]. These differences could have contributed to some of the preposition errors. Table 3 presents some error statistics for the adjuncts.

Overall, there are more deletions, but some prepositional complements are more prone to substitutions. For example, four-fifths of the preposition errors associated with "*university*" are subsitutions. Table 3 gives an example from the prepositional complement that suffers most often from each error type.

[2] Preposition errors in JLE are tagged under two categories, "PRP_LXC1" and "PRP_LXC2", which roughly correspond to this distinction. Insertion errors are also included in these categories, but they will be analyzed separately. See comments in §5.3.

[3] For example, the object particle "*wo*" normally does not correspond to any English word when marking a direct object in Japanese. However, in some contexts, it can necessitate a preposition in English, e.g., "*walk along a street*" from "*michi wo aruku*".

Type	Percent	Error	Percent	Example
Del	71.7%	[*to*→*null*]	42.2%	I want to **go**
		[*at*→*null*]	4.7%	[*to*→*null*]
		[*about*→*null*]	3.7%	Chicago
Sub	28.3%	[*with*→*to*]	3.0%	I'd like to
		[*to*→*for*]	2.3%	**exchange**
		[*to*→*with*]	1.9%	this [*with*→*to*]
				another one

Table 5. *Frequencies of deletions (Del) and substitutions (Sub) of prepositions in argument prepositional phrases, out of a total of 427. An example is drawn from the lexical head most frequently associated with a preposition error.*

Error Likelihood	
Context	Example
⟨**bump**,*into*⟩	A motorbike **bumped** [*into*→*to*] my car
⟨**ask**,*about*⟩	When the officer came to **ask** [*about*→*null*] the situation ...
⟨**graduate**,*from*⟩	He just **graduated** [*from*→*null*] university

Error Likelihood weighted w/ Frequency in General Domain	
Context	Example
⟨**look**,*at*⟩	She was **looking** [*at*→*null*] a monkey
⟨**ask**,*for*⟩	He **asked** [*for*→*null*] a table near the window
⟨**come**,*to*⟩	Last October , I **came** [*to*→*in*] Tokyo

Table 6. *The top table lists the verb-preposition (arguments) pairs with the highest error likelihoods (§4.1.1). The bottom table lists the top candidates when weighted with frequency in a general domain (§4.1.2).*

As in §4, to illustrate potential applications in grammar checkers and CALL systems, error likelihoods[4] are computed and weighted with frequencies in the general domain. Among the preposition-complement pairs with at least three appearances in the corpus, those with the highest likelihoods are shown in Table 4.

5.2. Arguments

Statistics on deletions and substitutions of prepositions in argument PPs are shown in Table 5. The proportion of deletions (71.7%) is substantially higher for arguments than for adjuncts, although the most prominent error, [*to*→*null*], is due mostly to the one verb "*go*", which alone is responsible for a third of the counts. Using the same procedure as for the adjuncts, the most error-prone verb-preposition pairs are listed in Table 6.

5.3. Insertion Errors

Preposition insertion errors belong to neither the adjunct nor argument categories, since no PP is actually needed. They typically occur when an adverb or adjective follows a verb, such as "*She asked him to bring the cat [null→to] home.*" or "*lives [null→in] next door*". The five most frequently inserted prepositions are *to*, *in*, *with*, *for* and *of*.

[4] Since argument and adjunct PPs cannot yet be distinguished automatically with high accuracy [21], the counts are normalized without this distinction, possibly leading to an overestimated denominator.

6. CONCLUSION

We presented an automatic method to estimate a model of grammatical errors from a corpus of transcribed non-native speech, and explained how it can be useful for CALL and grammar checking systems. It is a fine-grained model that conditions each error type on the most salient word in the context, as illustrated with examples from articles and prepositions. We plan to integrate this model into a system for article and/or preposition correction.

7. REFERENCES

[1] E. Izumi, K. Uchimoto, and H. Isahara, *Nihonjin 1200-nin no eigo supikingu kopasu*, ALC Press, Tokyo, Japan, 2004.

[2] W. L. Johnson, S. Marsella, and H. Vihjálmsson, "The darwars tactical language training system.," in *Proc. Interservice/Industry Training, Simulation and Educational Conference*, 2004.

[3] X. Yi, J. Gao, and W. B. Dolan, "A web-based english proofing system for english as a second language users," in *Proc. IJCNLP*, 2008.

[4] Y. Tsubota, T. Kawahara, and M. Dantsuji, "Practical use of english pronunciation system for japanese students in the call classroom," in *Proc. Interspeech*, 2004.

[5] G. Heidorn, "Intelligent writing assistance," in *Handbook of Natural Language Processing*, R. Dale, H. Moisi, and H. Somers, Eds. Marcel Dekker, Inc., 2000.

[6] E. Bender, D. Flickinger, S. Oepen, A. Walsh, and T. Baldwin, "Arboretum: Using a precision grammar for grammar checking in call," in *Proc. InSTIL/ICALL Symposium on Computer Assisted Learning*, 2004.

[7] S. Seneff, "Interactive computer aids for acquiring proficiency in mandarin," in *Proc. ISCSLP*, 2006.

[8] L. N. Michaud, F. McCoy, K, and C. A. Pennington, "An intelligent tutoring system for deaf learners of written english," in *Proc. 4th International ACM Conference on Assistive Technologies*, 2000.

[9] K. Knight and I. Chander, "Automated postediting of documents," in *Proc. AAAI*, 1994.

[10] M. Chodorow, J. R. Tetreault, and N.-R. Han, "Detection of grammatical errors involving prepositions," in *Proc. ACL-SIGSEM Workshop on Prepositions*, 2007.

[11] C. Brockett, W. B. Dolan, and M. Gamon, "Correcting esl errors using phrasal smt techniques," in *Proc. ACL*, 2006.

[12] J. Lee, "Automatic article restoration," in *Proc. HLT-NAACL Student Research Workshop*, 2004.

[13] J. Lee and S. Seneff, "Automatic generation of cloze items for prepositions," in *Proc. Interspeech*, 2007.

[14] K. Hakuta, "A case study of a japanese child learning english as a second language," *Language Learning*, vol. 26, pp. 321–351, 1976.

[15] D. Liu and J. Gleason, "Acquisition of the article the by nonnative speakers of english," *Studies in Second Language Acquisition*, vol. 24, pp. 1–26, 2002.

[16] Granger S., Dagneaux E., and F. Meunier, *The International Corpus of Learner English*, Presses universitaires de Louvain, Belgium, 2002.

[17] S. Gui and H. Yang, *Zhongguo Xuexizhe Yingyu Yuliaoku*, Shanghai Waiyu Jiaoyu Chubanshe, Shanghai, China, 2003.

[18] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proc. ANLP*, 1997.

[19] M. Collins, *Head-Driven Statistical Models for Natural Language Parsing*, Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, 1999.

[20] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman, New York, 1985.

[21] P. Merlo and E. Esteve Ferrer, "The notion of argument in prepositional phrase attachment," *Computational Linguistics*, vol. 32, no. 3, pp. 341–378, 2006.

[22] H. Suzuki and K. Toutanova, "Learning to predict case markers in japanese," in *Proc. ACL*, 2006.