Predicting the Spelling Difficulty of Words for Language Learners

Lisa Beinborn[◇§], **Torsten Zesch**[‡§], **Iryna Gurevych**[◇§]

◇ UKP Lab, Technische Universität Darmstadt
‡ Language Technology Lab, University of Duisburg-Essen
§ UKP Lab, German Institute for Educational Research
`http://www.ukp.tu-darmstadt.de`

Abstract

In many language learning scenarios, it is important to anticipate spelling errors. We model the spelling difficulty of words with new features that capture phonetic phenomena and are based on psycholinguistic findings. To train our model, we extract more than 140,000 spelling errors from three learner corpora covering English, German and Italian essays. The evaluation shows that our model predicts spelling difficulty with an accuracy of over 80% and yields a stable quality across corpora and languages. In addition, we provide a thorough error analysis that takes the native language of the learners into account and provides insights into cross-lingual transfer effects.

1 Introduction

The irregularities of spelling have been subject to debates for a long time in many languages. Spelling difficulties can lead to substantial problems in the literacy acquisition and to severe cases of dyslexia (Landerl et al., 1997). Learning orthographic patterns is even harder for foreign language learners because the phonetic inventory of their mother tongue might be quite different. Thus, they have to learn both the new sounds and their mapping to graphemes. English is a well-known example for a particularly inconsistent grapheme-to-phoneme mapping. For example, the sequence *ough* can be pronounced in six different ways as in *though*, *through*, *rough*, *cough*, *thought* and *bough*.[1]

In many language learning scenarios, it is important to be aware of the spelling difficulty of a word. In Beinborn et al. (2014), we analyzed that words with high spelling error probability lead to more difficult exercises. This indicates, that spelling difficulty should also be considered in exercise generation. In text simplification tasks (Specia et al., 2012), a quantification of spelling difficulty could lead to more focused, learner-oriented lexical simplification. Spelling problems are often influenced by cross-lingual transfer because learners apply patterns from their native language (Ringbom and Jarvis, 2009). Spelling errors can therefore be a good predictor for automatic natural language identification (Nicolai et al., 2013). Language teachers are not always aware of these processes because they are often not familiar with the native language of their learners. Automatic prediction methods for L1-specific spelling difficulties can lead to a better understanding of cross-lingual transfer and support the development of individualized exercises.

In this paper, we take an empirical approach and approximate spelling difficulty based on error frequencies in learner corpora. We extract more than 140,000 spelling errors by more than 85,000 learners from three learner corpora. Two corpora cover essays by learners of English and the third corpus contains learner essays in German and Italian. We then train an algorithmic

[1] IPA pronunciations from https://en.wiktionary.org: /ð o ʊ/, /θ ɹ u/, /ɹ ʌ f/, /k ɔ f/, /θ ɔ t/, and /b a ʊ/

model on this data to predict the spelling difficulty of a word based on common word difficulty features and newly developed features modeling phonetic difficulties. We make the extracted errors and the code for extraction and prediction publicly available.[2] Our evaluation results show that it is generally possible to predict the spelling difficulty of words. The performance remains stable across corpora and across languages. Common word features such as length and frequency already provide a reasonable approximation. However, if we aim at explaining the processes that cause different spelling errors depending on the L1 of the learner, phonetic features and the cognateness of words need to be taken into account.

2 Measuring Spelling Difficulty

Analyses of English spelling difficulties have a long tradition in pedagogical and psycholinguistic literature, but to the best of our knowledge the task of predicting spelling difficulty has not yet been tackled. In this section, we operationalize the analytical findings on spelling difficulty into features that can be derived automatically.

In general, three sources of spelling errors can be distinguished: i) errors caused by physical factors such as the distance between keys on the input device or omitted character repetitions, ii) errors caused by look-ahead and look-behind confusion (e.g. *puclic–public*, *gib–big*), iii) and errors caused by phonetic similarity of letters (e.g. vowel confusion *visable–visible*). Baba and Suzuki (2012) analyze spelling errors committed by English and Japanese learners using keystroke logs and find that the first two types are usually detected and self-corrected by the learner whereas phonetic problems remain unnoticed. In the learner corpora that we analyze, the learners were encouraged to review their essays thoroughly, so we focus on spelling errors that are usually not detected by learners.

In the following, we describe seven features that we implemented for spelling difficulty prediction: two word difficulty features (length and frequency) and five phonetic features (grapheme-to-phoneme ratio, phonetic density, character sequence probability, pronunciation difficulty and pronunciation clarity).

2.1 Word Difficulty Features

Many psycholinguistic studies have shown that frequency effects play an important role in language acquisition (Brysbaert and New, 2009). High-frequency words enable faster lexical access and should therefore be easier to memorize for language learners. For English, the word length is in principle a good approximation of word frequency because frequently used words tend to be rather short compared to more specific terms. Medero and Ostendorf (2009) and Culligan (2015) analyze vocabulary difficulty and find that short length and high frequency are good indicators for simple words. Both features are also highly relevant for spelling difficulty. Put simply, the probability of producing an error is increased by the number of characters that need to be typed. For frequent words, the probability that the learner has been exposed to this word is increased and therefore the spelling difficulty should be lower. We determine the length of a word by the number of characters and the frequency is represented by the unigram log-probability of the word in the Web1T corpus (Brants and Franz, 2006).

2.2 Phonetic Difficulty

In addition to the traditional features mentioned above, phonetic ambiguity has been intensely analyzed in the spelling research. Frith (1980) compares the spelling errors of good and poor readers and shows that good readers only produce phonetic misspellings whereas poor readers (which she called 'mildly dyslexic') often produce non-phonetic misspellings. Cook (1997) compares English spelling competence for L1 and L2 users. She confirms that the majority of spelling errors by all three groups (L1 children, L1 adults, L2 adults) are due to ambiguous sound–letter correspondences. Berkling et al. (2015b) study the interplay between graphemes and phonotactics in German in detail and developed a game to teach orthographic patterns

[2] https://www.ukp.tu-darmstadt.de/data/spelling-correction/spelling-difficulty-prediction

to children. Peereman et al. (2007) provide a very good overview of factors influencing word difficulty and also highlight the importance of consistent grapheme–phoneme correspondence. It thus seems justified to focus on the phonetic problems. The features described below try to approximate the relationship between graphemes and phonemes from various angles.

Orthographic Depth Rosa and Eskenazi (2011) analyze the influence of word complexity features on the vocabulary acquisition of L2 learners and show that words which follow a simple one-to-one mapping of graphemes to phonemes are considered to be easier than one-to-many or many-to-one mappings as in *knowledge*.[3] The orthographic depth can be expressed as the grapheme-to-phoneme ratio (the word length in characters divided by the number of phonemes). For English, we calculate the number of phonemes based on the phonetic representation in the Carnegie Mellon University Pronouncing Dictionary.[4] For Italian and German, a comparable pronunciation resource is not available. However, as the orthography of these two languages is more regular than for English, the pronunciation of a word can be approximated by rules. We use the grapheme-to-phoneme transcription of the text-to-speech synthesis software MaryTTS version 5.1.1 (Schröder and Trouvain, 2003) to determine the phonetic transcription for Italian and German. MaryTTS uses a mixture of resource-based and rule-based approaches. We will refer to transcriptions obtained from these resources as gold transcriptions.

Phonetic Density The phonetic density has also been proposed as a potential cause for spelling difficulty, but has not yet been studied extensively (Joshi and Aaron, 2013). It is calculated as the ratio of vowels to consonants. Both extremes—words with high density (e.g. *aerie*) and very low density (e.g. *strength*)—are likely to cause spelling problems.

Character Sequence Probability We assume, that the grapheme–phoneme correspondence of a word is less intuitive, if the word contains a rare sequence of characters (e.g. *gardener* vs *guarantee*). To approximate this, we build a language model of character trigrams that indicates the probability of a character sequence using the framework Berkeleylm version 1.1.2 (Pauls and Klein, 2011). The quality of a language model is usually measured as the perplexity, i.e. the ability of the model to deal with unseen data. The perplexity can often be improved by using more training data. However, in this scenario, the model is supposed to perform worse on unseen data because it should model human learners. In order to reflect the sparse knowledge of a language learner, the model is trained only on the 800–1000 most frequent words from each language. We refer to these words as the *Basic Vocabulary*.[5]

Pronunciation Difficulty Furthermore, we try to capture the assumption that a spelling error is more likely to occur if the grapheme–phoneme mapping is rare as in *Wednesday*. The sequence *ed* is more likely to be pronounced as in simple past verbs or as in *Sweden*. We approximate this by building a phonetic model using Phonetisaurus, a tool that is based on finite state transducers which map characters onto phonemes and can predict pronunciations for unseen words.[6] Analogous to the character-based language model, the phonetic model is also trained only on words from the *Basic Vocabulary* in order to reflect the knowledge of a language learner. Based on this scarce data, the phonetic model only learns the most frequent character-to-phoneme mappings and assigns higher phonetic scores to ambiguous letter sequences. We use this score as indicator for the pronunciation difficulty.

Pronunciation Clarity Even if the learner experiences low pronunciation difficulty, she

[3] grapheme length: 9, phoneme length: 5

[4] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[5] We use the following lists: en: http://ogden.basic-english.org, de: http://www.languagedaily.com/learn-german/vocabulary/common-german-words, it: https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Italian1000

[6] http://code.google.com/p/phonetisaurus

might still come up with a wrong pronunciation. For example, many learners are convinced that *recipe* should be pronounced /ɹ ɪ s a ɪ p/. To model the discrepancy between expected and true pronunciation, we calculate the Levenshtein distance between the produced pronunciation by the phonetic model and the gold transcription as pronunciation clarity.

3 Spelling Error Extraction

In order to evaluate the described model for predicting spelling difficulty, we need suitable data. For this purpose, we extract spelling errors from corpora of annotated learner essays. The corpora contain annotations for a wide range of errors including spelling, grammar, and style. As the corpora use different annotation formats, we implement an extraction pipeline to focus only on the spelling errors. We apply additional preprocesssing and compute the spelling error probability as an indicator for spelling difficulty.

3.1 Corpora

We use learner essays and error annotations from three corpora: EFC, FCE and Merlin. The first two contain essays by learners of English and the Merlin corpus contains essays by learners of German and Italian.[7] We describe them in more detail below.

EFC The EF-Cambridge Open Language Database (Geertzen et al., 2012) contains 549,326 short learner essays written by 84,997 learners from 138 nationalities. The essays have been submitted to *Englishtown*, the online school of *Education First*. 186,416 of these essays are annotated with corrections provided by teachers. We extract 167,713 annotations with the tag *SP* for spelling error.[8] To our knowledge, this is by far the biggest available corpus with spelling errors from language learners.

FCE The second corpus is part of the Cambridge Learner Corpus and consists of learner answers for the *First Certificate in English* (FCE) exam (Yannakoudakis et al., 2011). It contains 2,488 essays by 1,244 learners (each learner had to answer two tasks) from 16 nationalities. The essays have been corrected by official examiners. We extract 4,074 annotations with the tag *S* for spelling error.

Merlin The third corpus has been developed within the EU-project MERLIN (Boyd et al., 2014) and contains learner essays graded according to the Common European Reference Framework. The 813 Italian and the 1,033 German samples have been obtained as part of a test for the European language certificate (TELC). 752 of the German essays and 754 of the Italian essays were annotated with target hypotheses and error annotations by linguistic experts. We extract 2,525 annotations with the tag *O_graph* from the German essays and 2,446 from the Italian essays. Unfortunately, the correction of the errors can only be extracted, if the error annotation is properly aligned to the target hypotheses which is not always the case. We ignore the errors without available correction which reduces the set to 1,569 German and 1,761 Italian errors. In the following, we refer to the German subset as M-DE and the Italian subset as M-IT.

3.2 Error Extraction

As the annotation guidelines differed for the three corpora, we first need to apply additional pre-processing steps. In a second step, we aim at quantifying the spelling difficulty for each word by calculating the spelling error probability.

Pre-processing We remove all spelling errors that only mark a change from lowercase to uppercase (or vice versa) and numeric corrections (e.g. *1* is corrected to *one*) as these are rather related to stylistic conventions than to spelling. We lowercase all words, trim whitespaces and only keep words which occur in a word list and consist of at least three letters (to avoid abbreviations like *ms*, *pm*, *oz*).[9]

[7] It also contains essays by Czech learners, but this subset is significantly smaller than the ones for the other two languages and is therefore not used here.

[8] Some corrections have two different tags; we only extract those with a single *SP* tag.

[9] We use the word list package provided by Ubuntu for spell-checking: http://www.ubuntuupdates.org/package/core/lucid/main/base/\$PACKAGE, packages: wamerican, wngerman, wfrench

		EFC	**FCE**	**M-DE**	**M-IT**
Words	All	7,388,555	333,323	84,557	57,708
	Distinct	23,508	7,129	3,561	3,760
Spelling Errors	All	133,028	3,897	1,653	1,904
	Distinct	7,957	1,509	719	747
Ratio Errors/Words	Distinct	.34	.21	.20	.20

Table 1: Extracted words and spelling errors after pre-processing

Spelling Error Probability In this work, we take an empirical approach for quantifying spelling difficulty. A spelling error s is represented by a pair consisting of a misspelling e and the corresponding correction c. The error frequency f_e of a word w in the dataset D is then determined by the number of times it occurs as a correction of a spelling error independent of the actual misspelling. The number of spelling errors S_D in the dataset is determined by summing over the error frequencies of all words in the dataset. To quantify the distinct spelling errors, we count all words with $f_e \geq 1$ once.

$$s = (e, c) \tag{1}$$

$$f_e(w) = \sum_{s_i \in D} |w = c_i| \tag{2}$$

$$S_D = \sum_{w_i \in D} f_e(w_i) \tag{3}$$

The number of extracted words and errors are summarized in Table 1. It can be seen that the EFC corpus is significantly bigger than the other corpora. The spelling errors in the EFC corpus are spread over many words leading to a higher ratio of erroneous words over all words.

The pure error frequency of a word can be misleading, because frequently used words are more likely to occur as a spelling error independent of the spelling difficulty of the word. Instead, we calculate the spelling error probability for each word as the ratio of the error frequency over all occurrences of the word (including the erroneous occurrences).

$$p_{err}(w) = \frac{f_{err}(w)}{f(w)} \tag{4}$$

	Error Probability	
Corpus	**high**	**low**
EFC	*departmental*	*boy*
	spelt	*car*
	invincible	*crime*
FCE	*synthetic*	*weeks*
	millennium	*feel*
	mystery	*rainbow*
M-DE	*tschüss*	*damit*
	nächsten	*machen*
	beschäftigt	*gekauft*
M-IT	*messagio*	*rossi*
	lunedí	*questo*
	caffè	*tempo*

Table 2: Examples for high and low spelling error probability

The words are then ranked by their error probability to quantify spelling difficulty.[10] This is only a rough approximation that ignores other factors such as repetition errors and learner ability because detailed learner data was not available for all corpora. In future work, more elaborate measures of spelling difficulty could be analyzed (see for example Ehara et al. (2012)).

3.3 Training and Test Data

An inspection of the ranked probabilities indicates that the spelling difficulty of a word is a continuous variable which points to a regression problem. However, the number of spelling errors is too small to distinguish between a spelling error probability of 0.2 and 0.3, for example. Instead, we only focus on the extremes of the scale.

[10] In the case of tied error probability, the word with the higher error frequency is ranked higher. In the case of an error frequency of zero for both words, the word with the lower correct frequency is ranked higher.

		EFC	**FCE**	**M-DE**	**M-IT**
	Random Baseline	$.500^{**}$	$.500^{**}$	$.500^{**}$	$.500^{**}$
Individual Features	Orthographic Depth	$.482^{**}$	$.462^{**}$	$.427^{**}$	$.622^{**}$
	Phonetic Density	$.483^{**}$	$.349^{**}$	$.564^{**}$	$.508^{**}$
	Character Sequence Probability	$.706^{**}$	$.642^{**}$	.736	$.563^{**}$
	Pronunciation Clarity	$.635^{**}$	$.677^{**}$	.722	.683
	Pronunciation Difficulty	$.792^{**}$	$.792^{**}$	**.828**	.731
	Frequency	$.634^{**}$	$.742^{**}$	.778	.728
	Length	**.809**	**.827**	.747	**.769**
Combined	Length + Frequency + Pronunciation Diff.	.822	.832	**.828**	**.792**
	All Features	**.835**	**.847**	.814	**.778**

Table 3: Feature analysis for spelling difficulty using 10-fold cross-validation. The prediction results are expressed as accuracy. Significant differences compared to the result with all features are indicated with **(p<0.01).

The *n* highest ranked words are considered as samples for high spelling difficulty and the *n* lowest-ranked words form the class of words with low spelling difficulty. As additional constraint, the errors should have been committed by at least three learners in the EFC dataset and by two learners in the other corpora. For the EFC dataset, we extract 500 instances for each class, and for the FCE dataset 300 instances. 200 instances (100 per class) are used for testing in both cases and the remaining instances are used for training. We find an overlap of 52 words with high spelling error probability in both English corpora. As the Merlin corpus is significantly smaller, we only extract 100 instances per class for German and Italian. 140 instances are used for training and 60 for testing. Table 2 provides examples for high and low error probabilities.

4 Experiments & Results

The following experiments test whether it is possible to distinguish between words with high and low spelling error probability using the features described in Section 2. The models are trained with support vector machines as implemented in Weka (Hall et al., 2009). The features are extracted using the DKPro TC framework (Daxenberger et al., 2014).

4.1 Feature Analysis

In a first step, the predictive power of each feature is evaluated by performing ten-fold cross-validation on the training set. The results in the upper part of Table 3 are quite similar for the two English corpora. Around 80% of the test words are classified correctly and the most predictive features are the word length and the pronunciation difficulty. It should be noted, that the two features are correlated (Pearson's r: 0.67), but they provide different classifications for 131 of the 800 EFC instances in the cross-validation setting. The results for Italian are slightly worse than for English, but show the same pattern for the different features. For German, the pronunciation difficulty and frequency features perform slightly better than the length feature. The two features orthographic depth and phonetic density are not predictive for the spelling difficulty and only perform on chance level for all four datasets. We additionally train a model build on the three best performing features length, frequency, and pronunciation difficulty as well as one using all features. It can be seen that the results improve slightly compared to the individual features. Due to the rather small datasets and the correlation between the features, the differences between the best performing models are not significant.

In general, the accuracy results are comparable across languages (78–85%) indicating that it is possible to distinguish between words with high and low spelling error probability. In the following, we test whether the models can generalize to the unseen test data.

	EFC	**FCE**	**M-DE**	**M-IT**
Random	.500	.500	.500	.500
Len/Freq/Pron	**.840**	.865	.766	**.817**
All	**.840**	**.870**	**.800**	.815

Table 4: Spelling difficulty prediction on the test set for both corpora. The prediction results are expressed as accuracy.

4.2 Prediction Results

After these analyses, the two combined models are evaluated on the unseen test data. The results in Table 4 show that the models scale well to the test set and yield accuracy results that are slightly better than in the cross-validation setting. Again, the results of the two combined models are not found to be significantly different. There are two explanations for this. On the one hand, the test set is quite small (200 instances for English, 60 instances for German and Italian) which makes it difficult to measure significant differences. On the other hand, this result indicates that length, frequency and pronunciation difficulty are very predictive features for the spelling difficulty and the other features only have insignificant effects. The finding that longer words are more likely to produce misspellings is not surprising. For deeper psycholinguistic analyses it might be useful to balance the spelling data with respect to the word length. In such a scenario, phonetic aspects would presumably become more important. However, as we want to model the probability that a learner makes a spelling error, we need to take the length effect into account as an important indicator.

4.3 Cross-corpus comparison

The above results have shown that the prediction quality is very similar for the two English corpora. To analyze the robustness of the prediction approach, we compare the prediction quality across corpora by training on all instances of one corpus and testing on the instances of another. We also include the German and Italian corpora to this cross-corpus comparison to evaluate the language-dependence of spelling difficulty.

Train Corpus		**Test Corpus**			
		EFC	FCE	M-DE	M-IT
	# inst.	200	200	60	60
EFC	800	.840	.772	.703	.634
FCE	600	.764	.870	.767	.766
M-DE	140	.659	.829	.800	.796
M-IT	140	.397	.540	.780	.815

Table 5: Spelling difficulty prediction on the full set across corpora. The prediction results are expressed as accuracy. The number of instances is indicated in brackets for each dataset. The two classes are equally distributed.

The results in Table 5 show that the accuracy for cross-corpus prediction generally decreases compared to the previous results of in-corpus prediction (which are listed in the diagonal of the result matrix), but still remains clearly above chance level for English and German. In contrast, training on the Italian corpus leads to bad results for the two English corpora. It is interesting to note that a model trained on the German spelling errors performs better on the FCE words than a model trained on the English errors from the EFC corpus. The FCE and the Merlin corpus have been obtained from standardized language examinations whereas the EFC corpus rather aims at formative language training. In the second scenario, the learners are probably less prepared and less focused leading to more heterogeneous data which could explain the performance differences across corpora.

5 Error Analysis

For a more detailed analysis, we take a closer look at the mis-classifications for the EFC dataset. In a second step, we analyze spelling errors with respect to the L1 of the learners.

5.1 Misclassifications

The following words were classified as *high error probability*, but have a low error probability in the learner data: *references, ordinary, universal, updates, unrewarding, incentives, cologne, scarfs, speakers, remained, vocals.* It seems surprising that all those words should have a low error probability. A possible explanation could be that the words had been mentioned in the task description of the essays and are therefore frequently used and spelled correctly. Unfortunately, the task descriptions are not published along with the corpus and we cannot take this factor into account.

The words that were erroneously classified as words with a low spelling error probability are generally shorter: *icy, whisky, cried, curry, spelt, eight, runway, tattoo, daughter, farmers, discreet, eligible, diseases, typical, gallery, genre, mystery, arctic, starters, stretch, rhythm.* In several cases, we see phenomena for which features are available, e.g. a low vowel-consonant ratio in *stretch* and *rhythm*, an infrequent grapheme-to-phoneme mapping in *genre*, a low character sequence probability in *tattoo.* Unfortunately, these features seem to be overruled by the length feature.

In other examples, we observe phenomena that are specific to English and are not sufficiently covered by our features such as irregular morphology (*icy, spelt, cried*). This indicates that features which model language-specific phenomena might lead to further improvements.

5.2 Influence of the L1

As phonetic features have a strong influence on spelling difficulty, we assume that the L1 of the learners plays an important role. For example, *arctic* is misspelled as **artic*, *gallery* as **galery* and *mystery* and *typical* are spelled with *i* instead of *y*. These misspellings correspond to the correct stem of the respective word in Spanish, Italian and Portuguese. In the following, we thus have a closer look at the influence of the L1.

The EFC corpus comprises essays from a very heterogeneous group of learners, but 71% of the annotated essays are written by learners from five nationalities, namely Brazilian, Chinese, German, Mexican, and Russian. For comparative analyses, we also extracted the spelling errors specific to each of these five nationalities. Table 6 shows anecdotal examples of cross-lingual influence on spelling difficulties. For the word *attention*, it can be seen that the Russian learners are tempted to use an *a* as second vowel instead of an *e*. For the Brazilian and Mexican learners, on the other hand, the duplication of the *t* is more problematic because doubled plosive consonants do not occur in their L1.

L1-specific errors are often due to the existence of similar words—so-called cognates—in the native language of the learner. The word *departmental* is particularly difficult for Brazilian and Chinese learners. While most Brazilian learners erroneously insert an *a* due to the cognate *departamento*, none of the Chinese learners commits this error because a corresponding cognate does not exist. The Brazilian and Mexican misspellings of *hamburger* can also be explained with the cognateness to *hamburguesa* and *hambúrguer* respectively. A *g* followed by an *e* is pronounced as a fricative /x/ in Spanish and not as a plosive /g/. This indicates that the phonetic features should model the differences between the L1 and the L2 of the learner.

The word *engineer* provokes a large variety of misspellings. A common problem is the use of *e* as the second vowel, which could be explained with the spelling of the cognates (br: *engenheiro*, de: *Ingenieur*, ru: инженер transliterated as *inzhener*). However, the misspelling by the Mexican learners cannot be explained with cognateness because the Spanish spelling would be *ingeniero.* The spelling of *marmalade* with an *e* seems to be idiosyncratic to German learners.

The above analyses are only performed on an anecdotal basis and need to be backed up with more thorough experimental studies. The examples support the intuitive assumption that cognates are particularly prone to spelling errors due to the different orthographic and phonetic patterns in the L1 of the learner. The cognateness of words can be determined automatically using string similarity measures (Inkpen et al., 2005) or character-based machine translation (Beinborn et al., 2013).

The learners in the EFC corpus also differ in proficiency (e.g. German learners seem to be more advanced than Brazilian learners) which might also have an influence on the spelling error probability of words. However, it is complicated to disentangle the influence of the L1 and

Correct	Brazilian	Mexican	Chinese	Russian	German
attention	*atention(27)*	*atention (13)*	*attaention (1)*	*attantion (5)*	-
	attencion (10)	*attencion(1)*	*atttention (1)*	*atantion (1)*	-
	atencion (3)	*attentio (1)*	-	*atention (1)*	-
departmental	*departament (10)*	*department (1)*	*deparment (2)*	-	-
	departamente (1)	-	*deparmental (1)*	-	-
	departaments (1)	-	*deprtment (1)*	-	-
hamburger	*hamburguer (2)*	*hamburguer (2)*	*hamburg*	-	-
	hamburguers (2)	-	*hamburgs (1)*	-	-
engineer	*engeneer (17)*	*enginner (25)*	*engneer (5)*	*engeneer (14)*	*ingeneur (2)*
	ingineer (2)	*engeneer (8)*	*engeneer (4)*	*engeener (3)*	*engeneer (2)*
	ingener (2)	*engenier (4)*	*enginner (3)*	*ingener (2)*	*ingeneer (2)*
marmalade	-	-	-	-	*marmelade (3)*

Table 6: Most frequent misspellings for selected examples

of the L2 proficiency based on the current data and we leave this analysis to future work.

6 Related work

In section 2, we already discussed psycholinguistic analyses of spelling difficulty. In natural language processing, related work in the field of spelling has focused on error correction (Ng et al., 2013; Ng et al., 2014). For finding the right correction, Deorowicz and Ciura (2005) analyze probable causes for spelling errors. They identify three types of causes (mistyping, misspelling and vocabulary incompetence) and model them using substitution rules. Toutanova and Moore (2002) use the similarity of pronunciations to pick the best correction for an error resulting in an improvement over state-of-the-art spellcheckers. Boyd (2009) build on their work but model the pronunciation of non-native speakers, leading to slight improvements in the pronunciation-based model. Modeling the spelling difficulty of words could also have a positive effect on spelling correction because spelling errors would be easier to anticipate.

Another important line of research is the development of spelling exercises. A popular recent example is the game Phontasia (Berkling et al., 2015a). It has been developed for L1 learners but could probably also be used for L2 learners. In this case, the findings on cross-lingual transfer could be integrated to account for the special phenomena occurring with L2 learners.

7 Conclusions

We have extracted spelling errors from three different learner corpora and calculated the spelling error probability for each word. We analyzed the concept of spelling difficulty and implemented common word difficulty features and new phonetic features to model it. Our prediction experiments reveal that the length and frequency features are a good approximation for spelling difficulty, but they do not capture phonetic phenomena. The newly developed feature for pronunciation difficulty can close this gap and complement the word difficulty features for spelling difficulty prediction.

We conclude that the spelling error probability of a word can be predicted to a certain extent. The prediction results are stable across corpora and can even be used across languages. A detailed error analysis indicates that further improvements could be reached by modeling language-specific features (e.g. morphology) and by taking the L1 of the learner into account. We make the spelling errors and our code publicly available to enable further research on spelling phenomena and hope that it will lead to new insights into the processes underlying foreign language learning.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Klaus Tschira Foundation under project No. 00.133.2008.

References

Yukino Baba and Hisami Suzuki. 2012. How are spelling errors generated and corrected? A study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 373–377. Association for Computational Linguistics.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-based Machine Translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 883–891. Asian Federation of Natural Language Processing.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–529.

Kay Berkling, Nadine Pflaumer, and Alexei Coyplove. 2015a. Phontasia—a game for training german orthography. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, pages 1874–1875.

Kay Berkling, Nadine Pflaumer, and Rémi Lavalley. 2015b. German Phonics Game using Speech Synthesis-A Longitudinal Study about the Effect on Orthography Skills. In *Proceedings of the Workshop on Spoken Language Technology for Education (SLaTE)*, pages 168–172.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Adriane Boyd. 2009. Pronunciation Modeling in Spelling Correction for Writers of English As a Foreign Language. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop and Doctoral Consortium*, pages 31–36. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1.1. *Linguistic Data Consortium*.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4):977–990.

Vivian Cook. 1997. L2 users and English spelling. *Journal of Multilingual and Multicultural Development*, 18(6):474–488.

Brent Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland, June. Association for Computational Linguistics.

Sebastian Deorowicz and Marcin G Ciura. 2005. Correcting spelling errors by modelling their causes. *International Journal of Applied Mathematics and Computer Science*, 15(2):275.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 799–814.

Uta Frith. 1980. Unexpected spelling problems. *Cognitive Processes in Spelling.*

Jeroen Geertzen, Dora Alexopoulou, and Anna Korhonen. 2012. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In Ryan T. Miller, editor, *Selected Proceedings of the 2012 Second Language Research Forum.* MA: Cascadilla Proceedings Project.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18.

Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 251–257.

R.M. Joshi and P.G. Aaron. 2013. *Handbook of Orthography and Literacy.* Taylor & Francis.

K Landerl, H Wimmer, and U Frith. 1997. The impact of orthographic consistency on dyslexia: a German-English comparison. *Cognition*, 63(3):315–34, June.

Julie Medero and Mari Ostendorf. 2009. Analysis of Vocabulary Difficulty using Wiktionary. *Proceedings of the 2nd Workshop on Speech and Language Technology in Education.*

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 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.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and Misspelling Features for Natural Language Identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145.

Adam Pauls and Dan Klein. 2011. Faster and Smaller N-Gram Language Models. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.*

Ronald Peereman, Bernard Lété, and Liliane Sprenger-Charolles. 2007. Manulex-infra: Distributional characteristics of grapheme—phoneme mappings, and infralexical and lexical units in child-directed written material. *Behavior Research Methods*, 39(3):579–589, August.

Håkan Ringbom and Scott Jarvis. 2009. The importance of cross-linguistic similarity in foreign language learning. In Michael H Long and Catherine J Doughty, editors, *The Handbook of Language Teaching*, chapter 7, pages 106–118. John Wiley & Sons.

Kevin Dela Rosa and Maxine Eskenazi. 2011. Effect of word complexity on l2 vocabulary learning. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–80. Association for Computational Linguistics.

Marc Schröder and Jürgen Trouvain. 2003. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 347–355. Association for Computational Linguistics.

Kristina Toutanova and Robert C Moore. 2002. Pronunciation modeling for improved spelling correction. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*