What is the difference between common spelling mistakes from German and Chinese native speakers in their second language English?

Annika Österdiekhoff

University of Duisburg-Essen, Germany
annika.oesterdiekhoff@stud.uni-due.de

Abstract

This paper is about the difference between common spelling mistakes from German and Chinese native speakers in their second language English. We checked the spelling from 247 essays from German and Chinese native speakers and figure out the various error categories. We evaluate the distribution of the error categories and consider the words with the highest probability to be misspelled.

1 Introduction

Nowadays people from all over the world learn English as their second language. An interesting question is, if the typical spelling mistakes differ by different native languages. A typical example is the misspelled word "informations". In German the word "information" has got a plural, but in English the plural is "information". This can be a case where German native speakers produce this error rather than Chinese native speakers.

Selinker (1972) defines the so-called "interlanguage". The interlanguage is the language a person speaks and writes when he learns a second language. This means he speaks and writes a second language which does not conform with the native language of the second language. Selinker (1972) characterise that the interlanguage depends on different aspects. One important aspect is the native language. Another aspect is the term "L1 transfer" (Weinreich, 2010), which describes the transfer of structures of the first language to the second language. Therefore, we think that there could be a difference in the common spelling mistakes. In addition to that the German and Chinese languages are very different which is the reason why we choose to compare between them. The Chinese language has got completely different characters than German and English, who use the same characters. Therefore, it can be, that German native speakers have other spelling mistakes due to they knowing a similar word in German.

We want to check if there is a difference between common spelling mistakes from German and Chinese native speakers in their second language English and in which way the difference is expressed. Therefore, we report previous work on this topic in Section two. In Section three we firstly describe our data set, and how we programmed and annotated our experiment. Our results are presented in Section four. We evaluate the distribution of the error categories, the most common misspelled words and the annotation of the error categories. After that we list some critical points of our research in Section five and sum up with our conclusion in Section six.

We hope our research can be used in language schools or other educational institutions to prevent typical errors for this group of native speakers.

2 Related Work

Our data set is based on the Native Language Identification Task (abbreviated as NLI) (Malmasi et al., 2017), which is a shared task to identify the native language of a writer, who writes in his second language English. The data is described in detail in Section 3.1.

One can see there is a fascinating connection between errors in texts and the native language of the writer. So Koppel et al. (2005) analyse how you can identify the author's native language by the errors of a text. They do not only focus at spelling mistakes, but also at syntax errors, neologism and part-of-speech bigrams and listed these errors very well in their paper. Other teams like Gebre et al. (2013) use also the NLI corpus (Malmasi et al., 2017) and figure out individual features for every language. In our case the features for German and Chinese native speakers are interesting.

But there are lots of other possibilities to use the data set. Like Blanchard et al. (2013) one can use it for the identification of the native language, but also for grammatical error correction or automatic essay scoring. Thus, one can use these things for educational applications. The team described very detailed the corpus of the native language identification task and the automatic essay scoring.

3 Experiment

To present our results we first have to describe our experimental set-up. Therefore, we first describe our used data set, then how we automatically select the spelling errors in the essays and in the end our used annotation scheme.

3.1 The data

We worked with the data from the NLI Shared Task (Malmasi et al., 2017). The task has got three different data bases. Firstly the native language identification by written essays, secondly the identification by spoken responses and thirdly by both of them. In our experimental setup we only use the written essays, because spoken responses have no spelling mistakes. The essays have got eight prompts, so we look up by which prompt the number of German native speakers and the number of Chinese native speakers are mostly the same (Blanchard et al., 2013). We choose prompt four, which has got 140 essays from German native speakers and 127 Chinese native speakers. We use two dictionaries to check the spelling. The dictionaries are from Hunspell[1] and from Jazzy[2].

3.2 Automatically select spelling errors in the essays

To write a program which select all spelling errors from the essays we first read out the file with the labels for the essays. Every entry of the file has got a test-taker-id with his speech-prompt, essay-prompt and his native language L1. We select the test-taker-id, where the essay-prompt is four and the L1 is GER (for German) or CHI (for Chinese). After that we run through every word of the selected essays. We use the Jazzy-Checker from DKPro (de Castilho and Gurevych, 2014), which is an annotator, who uses a dictionary to decide that the word is spelled correctly or not. This Jazzy-Checker runs over all essays and generates a list of all spelling errors for each language and each dictionary. For the evaluation we count all words from the essays for every language to put the number of spelling errors in relation.

3.3 Annotation

For the annotation we first take a look at the different outputs from the two dictionaries. We choose the Hunspell dictionary because it includes significantly more proper names. Thus, we have got less selected "mistakes" which are no mistakes. Then we want to classify the spelling errors. We refer to Koppel et al. (2005), who defined eight categories for spelling mistakes. The list is directly taken form Koppel et al. (2005):

- Repeated letter (e.g. $remmit$ instead of $remit$)
- Double letter appears only once (e.g. $comit$ instead of $commit$)
- Letter α instead of letter β (e.g. $firsd$ instead of $first$)
- Letter inversion (e.g. $fisrt$ instead of $first$)
- Inserted letter (e.g. $friegnd$ instead of $friend$)
- Missing letter (e.g. $frend$ instead of $friend$)
- Conflated words (e.g. $stucktogether$ [instead of $stuck\ together$])
- Abbreviations ([e.g. $21st$ instead of $twenty-first$])

We add three more categories:

- cannot be identified
- different English
- no mistake

The first new category is "cannot be identified". To assign a misspelled word to an error category, we have to assume which word the writer wants to say. The word which we assume is called “target hypothesis" in the automated error analysis (Reznicek et al., 2013). Consequently, the category “cannot be identified" is used for words where we do not find out which word the writer of the essay wants to say which means we find no clear target hypothesis. One example is the word "ont" which can mean onto or wont or something completely different.

The second category is "different English" which is necessary because the dictionary only

[1] https://github.com/wooorm/dictionaries/blob/master/dictionaries/en-US/index.dic

[2] https://sourceforge.net/projects/jazzy/files/Dictionaries/English/english.0.zip/download

contains the American English words and not the spelling of British English words.

In addition to that we have the category "no mistake". Because the dictionary does not know all proper names like the word "MSN", the web portal from Microsoft. So words like this are no spelling error.

4 Evaluation

Now we evaluate our approach to investigate the common spelling mistakes between German and Chinese native speakers in their second language English. For this we present the results of the distribution of the error categories as well as the differences in the number of spelling mistakes for the two native speaker groups. In addition to that we present the most common misspelled words and some special abnormalities between the different native languages as well as the evaluation of our annotation.

4.1 Results of the distribution of the error categories

The percentage of every error category for the native languages German and Chinese are shown in Figure 1.

Most of the remaining categories have nearly the same percentage. One interesting fact is, that the German native speakers use significant more British English words instead of American English as the Chinese native speakers. The percentage point difference is 6.42%. This could be because in German schools the teachers teach British English. A second difference is in the category "repeated letter" which is about 5% higher for the German native speakers. A reason for this difference can be the different native languages. The German language uses the same characters as the English language and has got double consonants. The Chinese language is based on different characters, where one character is a word or syllable and not a single letter (McBride-Chang and Ho, 2000). Therefore, the Chinese native speakers maybe have got less trouble to use repeated and single letters in the right way in English because they do not know repeated letters in their language. We do not know how to explain the difference of ten percent in the category "missing letter" between the German and Chinese native speakers. The distribution of the other categories is mostly the same for Chinese and German native speakers so all in all we can say that there is no significantly difference between the Chinese and German native speakers.

4.2 Evaluation of the number of spelling mistakes

	Ger.	Chi.
number of spelling mistakes	871	1108
number of misspelled words	810 (1.70%)	1054 (2.59%)
total number of words	47623	40748

Table 1: number of spelling mistakes and misspelled words in relation to the total number of words from German native speakers (here abbreviated with Ger.) and Chinese native speakers (here abbreviated with Chi.)

The Table 1 shows the number of spelling mistakes and number of misspelled words for the German and Chinese native speakers. We differ between spelling mistakes and misspelled words because in one word can be more than one spelling error, so it is possible that some words have got more than one category. An example is the word "diferrent" which we assume should be the word "different". This is in the first category "repeated letter" because the "r" is written twice instead of once and also in the category "double letter appears only once", because there is only one "f" instead of two. Thus, we have got overall more mistakes than misspelled words.

The percentage difference between the number of misspelled words is 0.89%. This means the Chinese native speaker have got one and a half times more misspelled words than the German native speakers. This is a significant difference between the two native languages.

Important to say is that the last two categories as described in the Figure 1 above are not included in the calculation in Table 1 because these are no real mistake.

4.3 Most common misspelled words

We figure out the words that have been misspelled the most. These words and the number of how often they are misspelled, are almost the same for the Chinese and German native speakers. The word which is mostly misspelled is the word "knowl-

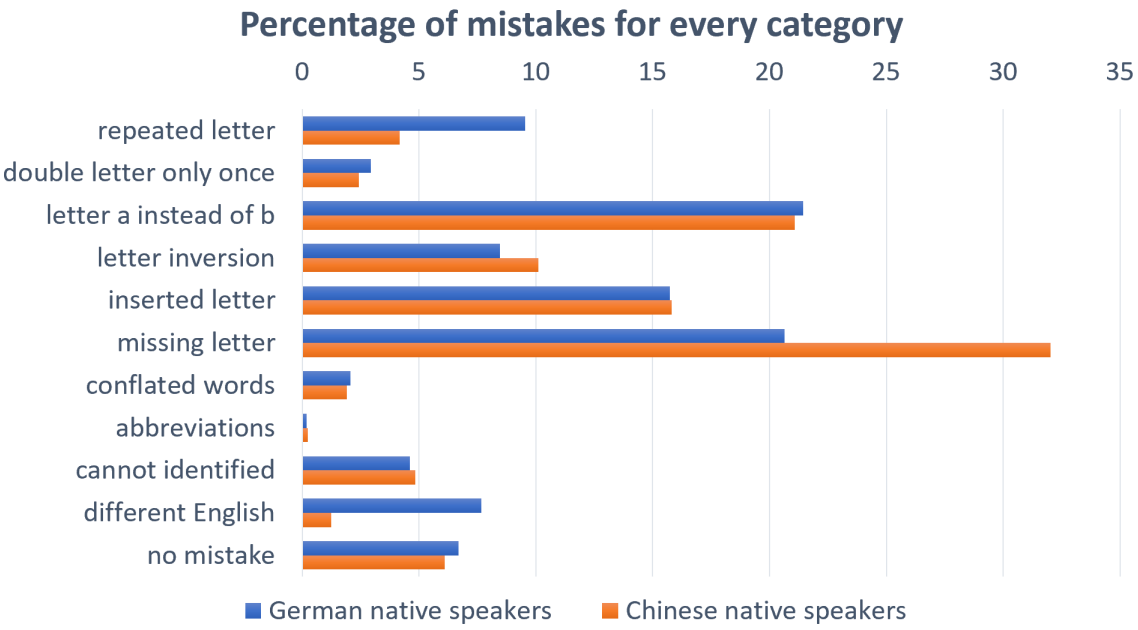

Figure 1: Percentage of spelling mistakes for every error category for German and Chines native speakers

	GER	no.	CHI	no.
1	knowlegde	30	knowlege	24
2	knowlege	15	knowledges	18
3	knownledge	9	konwledge/ knoledge	5

Table 2: The three most frequently occurring misspelled forms of the word knowledge for the German and Chinese native speakers. The number of the occurrence is given.

edge". For the Chinese it is incorrect 75 times in contrast to 77 times for the German. We list the three most frequently misspelled forms of the word "knowledge" in Table 2. We see that the words differ instead of the most misspelled version of knowledge of the Chinese native speakers with the second most misspelled version of the German native speakers. The number of misspelling forms of knowledge has got a big gap to the other numbers of misspelled words. The second mostly misspelled word is a form of "specialise". Forms are special, specialist, specialised, specialisation, and so on. The number again is nearly the same for both native speaker groups, like 34 times for the Chinese native speakers and 35 times for the German native speakers.

Interesting is that the number of the correctly written word knowledge is also pretty similar. The word is correctly used 758 times of the Chinese native speakers and 747 times of the German native speakers. This means that the two native groups uses the word equally often and misspell the words equally often. The same occurs for the correct versions of forms of specialised. These are also used equally often, namely 561 times of the Chinese native speakers and 584 times of the German native speakers. Our example from the beginning with the "s" in the word "information" pop up 8 times for the German native speakers, but zero times for the Chinese native speakers. All in all the difference between the most common misspelled words for Chinese and German native speakers is not big.

4.4 Special abnormalities between the different native languages

We detect some abnormalities between the different native languages during the annotation. Some important fact is that German native speakers use lots of words with a hyphen. This fact was also noticed by (Gebre et al., 2013) who has got one feature for the German native speakers for hyphen. An additional interesting fact is, that almost all the upper and lower case mistakes are words, which are a country, language or an inhabitant. This is due to the fact that in English most of the words are written in lower case but geographical features are written in upper case (Woods, 2017).

But one very important fact is that the German native speakers write more words, but have got fewer mistakes and a significant better scor-

ing (Blanchard et al., 2013). The German native speakers essays have got fundamental less misspelled words and are significantly better rated than the essays of the Chinese native speakers. For example from 1100 essays written by the German native speakers 673 have got the rating "high" in contrast to 275 "high" rated essays from the Chinese native speakers. This interesting discovery should be tested in the future.

4.5 Evaluation of the annotation with the help of a second annotator

		annotator two	
		0	1
annotator one	0	728	17
	1	20	60

Table 3: Confusion matrix between annotator one and annotator two about the annotation of the misspelled words

To check our classification of the error categories, we annotate the mistakes with a second annotator. The second annotator annotates the first 75 misspelled words of each language. For the evaluation we present the confusion matrix between these two annotators for all error categories together in Table 3. The confusion matrix has got four cases: (1) both annotators say that the word is not in this error category; (2) the case that one annotator say this word is in this error category and the other say this word is not in this error category; (3) the second case in reverse for the annotators; (4) and the case that both annotators say that the word is in this error category. We see that the the two annotator usually agree that the word is in the error category or that the word is not in the error category.

We calculate the cohens kappa of this total confusion matrix to evaluate the agreement of the two annotator (Cohen, 1960). It involves the fact that the two annotator agree by chance. The cohens kappa of this confusion matrix is 0.74. Landis and Koch (1977) define a interpretation of the cohens kappa. A value of 0.74 describes a "substantial" annotator agreement and is the second best description of agreement. To sum up the annotator agreement is very good and the annotator annotates the error categories very similar.

5 Limitations

We should look differentiated on the results. There are possible reasons why the results might be wrong. One reason can be when the annotators assume for a misspelled word different target hypotheses, therefore different correct forms of the word. So maybe some number of categories should be higher or lower. Another important aspect is the situation in which the essays have been written. The essays from our data are especially for the native language identification shared task. Therefore, we can assume that lots of essays are not written with exam feeling or with the intention to get a good grade. If we would use essays from exams or something similar the writers maybe had been more carefully about spelling mistakes.

Another topic to think about is that we have the essay scoring only for all 1100 essays of the data set. So it can be, that the essays in prompt four are rated completely different from the others and we cannot differ that. So our assumption that the essays of the German native speakers is rated much better than the Chinese rated speakers could be wrong. Also, the thesis that the number of spelling mistakes can influence the scoring should be checked in detail.

6 Conclusion

All in all we can say that there is no big difference in the type of spelling mistakes. So Chinese native speakers leave out a letter more often, but on the other hand German native speakers falsely repeat letters more often. In general most of the distribution among the error categories is nearly the same. But a difference between the German and Chinese native speakers is the total number of misspelled words and thereby the scoring of the essays.

It also could be interesting to analyse the other prompts of the data sets and see these results. Maybe they are different from the prompt four and we accidentally chose a prompt with untypical results.

In the future it would be interesting to know the analysis results for different languages. Maybe there is another difference or it is similar to the analysis of the native languages German and Chinese. Another good question is to analyse the impact from spelling mistakes to the scoring. We thought the impact would be much greater than it seems in our analysis. It would be a good idea to research this in more detail.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series* 2013(2):i–15.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. pages 1–11.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. pages 216–223.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *KDD*. volume 5, pages 624–628.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* pages 363–374.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A Report on the 2017 Native Language Identification Shared Task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Copenhagen, Denmark.

Catherine McBride-Chang and Connie Suk-Han Ho. 2000. Developmental issues in chinese children's character acquisition. *Journal of Educational Psychology* 92(1):50.

Marc Reznicek, Anke Ludeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the falko corpus. *Automatic treatment and analysis of learner corpus data* 59:101–123.

Larry Selinker. 1972. Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching* 10(1-4):209–232.

Uriel Weinreich. 2010. *Languages in contact: Findings and problems*. 1. Walter de Gruyter.

Geraldine Woods. 2017. *English grammar for dummies*. John Wiley & Sons.