Misspellings in Responses to Listening Comprehension Questions: Prospects for Scoring based on Phonetic Normalization

Heike da Silva Cardoso[†] and **Magdalena Wolska**[*]
[†]Department of Linguistics [*]LEAD Graduate School
Eberhard Karls Universität Tübingen, Tübingen, Germany
{hcardoso,magdalena.wolska}@uni-tuebingen.de

Abstract

Automated scoring systems which evaluate content require robust ways of dealing with form errors. The work presented in this paper is set in the context of scoring learners' responses to listening comprehension items included in a placement test of German as a foreign language. Based on a corpus of over 3000 responses to 17 questions, by test takers of different language proficiencies, we perform a quantitative analysis of the diversity in misspellings. We evaluate the performance of an off-the-shelf open source spell-checker on our data showing that around 45% of the reported non-word errors are not correctly accounted for, that is, they are either falsely identified as misspelt or the spell-checker is unable to identify the intended word.

We propose to address misspellings in computer-based scoring of constructed response items by means of phonetic normalization. Learner responses transcribed into Soundex codes and into two encodings borrowed from historical linguistics (ASJP and Dolgopolsky's sound classes) are compared to transcribed reference answers using string distance measures. We show that reliable correlation with teachers' scores can be obtained, however, similarity thresholds are item-specific.

1 Introduction

Form errors are the type of noise in linguistic data that can interfere with computational language analysis already at the preprocessing stage. Form errors in writing range from basic mechanics errors, such as capitalization or punctuation errors, through spelling and word-formation errors (which in many cases cannot be clearly differentiated), up to sentence structure, syntactic, errors. In this paper we address one class of form errors, non-word misspellings, in the context of a semantics-oriented task: assessment of constructed responses to German as a Foreign Language listening comprehension questions.

In the task of content scoring, misspellings introduce obvious noise. A recently proposed method of addressing the spelling problem in *automated* scoring involves phonetic normalization based on Soundex, a coarse-granularity sound-based coding. Shedeed (2011) used Soundex in a system for scoring short answers in Arabic. Hahn et al. (2013) used an analogous method for German and showed that a bag-of-Soundex model outperforms other models on unseen data at the accuracy over 85%.

The work presented here has been motivated by a different approach to content scoring: *computer-assisted* scoring. In the context of a real-world task, instead of automatically assigning scores we group responses that are likely to be graded with the same scores with the goal of streamlining manual scoring (see (Wolska et al., 2014)). Identifying responses that are similar at the appropriate level of abstraction is thus crucial here. In the study presented in this paper, we evaluate the prospects for using phonetic string encodings based on sound classes derived in historical linguistics as a preprocessing step for this task.

In historical and comparative linguistics sound classes are used, among others to detect cognates, identify relatedness among languages, or detect or explain changes in sound patterns. Phonetic encoding in this case is a normalization step which serves to make languages comparable. In our case, phonetic normalization of type-written responses to *listening* comprehension items is motivated by the fact that students, especially those of lower

*Corresponding author

Heike Da Silva Cardoso and Magdalena Wolska 2015. Misspellings in responses to listening comprehension questions: Prospects for scoring based on phonetic normalization. *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015.* NEALT Proceedings Series 26 / Linköping Electronic Conference Proceedings 114: 1–10.

proficiency, tend to misspell words to some extent in systematic ways, for instance, related to the properties of their mother-tongue (orthography rules or phonological differences between the mother-tongue and the target language).

Based on a corpus of learner responses to listening comprehension items, in this paper we answer the following questions:

- What is the extent of the misspellings problem in learner responses to German listening comprehension questions?
- How diverse are misspellings, that is, to what extent they diverge from target hypotheses?
- To what extent an off-the-shelf spell-checking tool can "solve" the problem?
- Does grouping responses based on phonetic normalization account for teacher's response scores?

In the context of the last question, we test two linguistically-motivated phonetic encodings of different granularity: ASJPcode (Wichmann et al., 2013) and Dolgopolsky's classes (Dolgopolsky, 1986). These are compared to Soundex encoding (Russell, 1918 1922), a practically-motivated indexing method, which, as mentioned earlier, had been previously proposed as a preprocessing step in content scoring. We hypothesize that normalization based on the linguistically-motivated systems should yield response groups that better reflect the assigned scores than grouping based on Soundex encoding.

2 Related Work

Research into misspellings in learner language has been predominantly addressing English as the target (see, for instance, (Flor and Futagi, 2012) for a recent overview). Analogous lines of work based on digital corpora has been emerging for German as a Foreign Language. Rimrott and Heift (2008) analyzed the performance of MS Word spell-checker on learner German and found that around 20% of misspellings were undetected. For single-error words, in over 40% of the cases the correct word was not in the suggestion list whereas for multiple-error words in about 80% of the cases the spell-checker failed to provide a correction. In a further study, Heift and Rimrott (2008) found that in CALL activities students are influenced by a word's position in the list of suggestions when they select an alternative spelling. Clearly, with incorrect top-level suggestions, only more errors are introduced.

Corpus-based studies into low-level form errors in German learner writing are sparse. Boyd (2010) created a corpus of online workbook exercises and essays submitted of by American students learning German and built a subcorpus of around 1200 non-word spelling errors found in this data. The most prominent error annotated German learner corpus is Falko (Reznicek et al., 2013) and it also includes annotations of target hypotheses for misspellings. Juozulynas (2013) analyzed around 350 German essays written by American college students and found that around 15% of the identified errors were spelling errors. Analysis of accuracy of robust automated correction was not performed in these studies.

To our knowledge, the only prior work in which explicit phonetic normalization is employed in content scoring is the previously mentioned work by Shedeed (2011) and the subsequent study by Hahn et al. (2013). In both cases Soundex coding is used.

3 Listening Comprehension Corpus

Data collection In this study we used responses to listening comprehension (LC) items collected during placement tests for language courses (four cohorts of students) administered by the Saarland University's International Office centre for Teaching German as a Foreign Language. The tests consisted of three parts: grammar, C-Test, and listening comprehension. The listening part consisted of three audio stimuli of increasing difficulty in terms of linguistic properties and speech tempo. The stimuli were accompanied with up to 11 constructed response questions each. For each question the teachers provided one or more correct reference answers.

The tests were developed by an experienced teacher of the language centre and conducted using a web-based platform. Students' responses, preprocessed as outlined below, were scored manually – for the most part one teacher, head of the centre – also using a web-based platform. Responses were graded on a [0,1] or [0,2] scale; half points were used for partial credits. Approximately 600 students of various proficiencies and mother tongues participated in the tests.

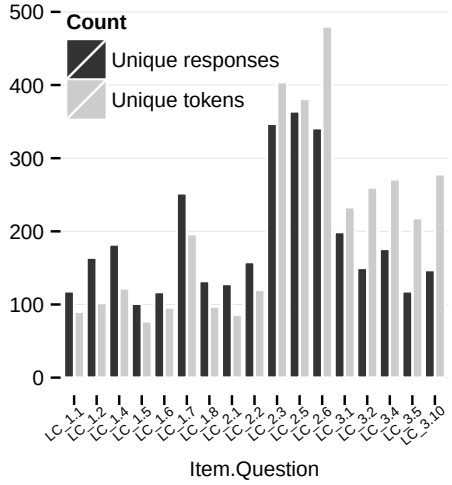

Figure 1: Number of unique responses and unique tokens per question

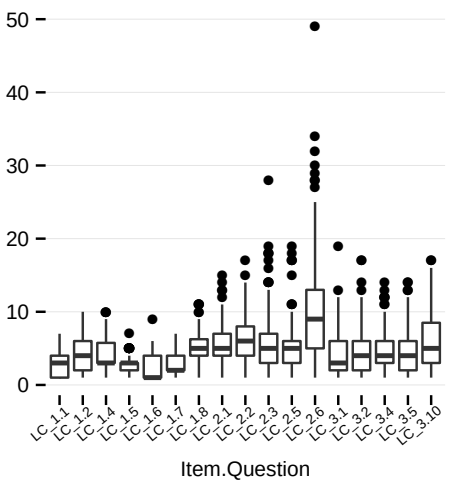

Figure 2: Response lengths, in tokens, per question

Variable	N
Verbatim responses	7208
Verbatim unique	3794
Preprocessed unique	3146
Tokens	16298
Token types	2429

Table 1: Descriptive corpus information

Preprocessing Certain minor form errors, such as wrong capitalization or irregular punctuation, are irrelevant while assessing comprehension. We exploit this in a scoring platform to reduce the set of responses to score by normalizing spurious writing mechanics differences which are not considered score-affecting in assessing comprehension. This includes lower-casing and removing clause- and sentence-final punctuation. In order to avoid differences in edit distance due to diacritics use, we also transcribe umlaut characters, using the standard convention, with their underlying vowel followed by 'e' ('ö' as 'oe', 'ü' as 'ue', etc.). Preprocessing reduces the set of responses which teachers need to score by more than 50% for some items. For this study we use responses scored in the preprocessed form. For the analysis presented in this paper we use a subset of the scored preprocessed responses selected as summarized below.

The corpus Since the number of responses differs from question to question (at least partially due to different language proficiencies of the test-takers; low-proficiency test-takers are not capable of responding to questions to the more difficult audio prompts) and for some questions it is low (only 29 responses to one of the questions after preprocessing) for the analyses presented in this paper, we selected only those questions to which we have at least 100 unique preprocessed responses. We moreover excluded questions which elicited unordered multi-part responses, that is, questions of the type "Name 3 ..." or "What are ...? (2 items)". Our complete data set consists of responses to 17 questions which elicited single-part responses and each response has been scored at 0, 0.5, or 1 points.

Table 1 shows basic descriptive information about the corpus. The number of verbatim responses is the total number of responses to the 17 questions before preprocessing. "Verbatim unique" is the number of token-identical verbatim responses collapsed to one observation. "Preprocessed unique" is the number of token-identical (unique) responses after preprocessing as described in the previous paragraph. "Tokens" and "Token types" are, respectively, the number of all tokens and unique tokens (types) in the preprocessed responses.

In the remainder of this paper, we refer to the set of preprocessed unique responses. Figure 1 shows the distribution of responses and unique tokens per question for the three items (LC_1, LC_2, LC_3). Figure 2 shows the distribution of response lengths per question. There are more unique responses to the more difficult items, LC_2 and LC_3, and the responses to those items are longer and more di-

LC_1.1	LC_1.6	LC_3.1
frankreisch	austereich	giespallampe
frankrich	austerreich	energiespaerlaempe
frankriech	oestereicht	energysparen
frankrreisch	oeustreich	energiesparenlampen
frankrreit	ostreich	energiesparlampel
franzoezisch	oesterreisch	energiesparer
franzuezisch	oesttereich	energiespannlampe
freinkreich	oeustreich	energisparelampen
frienkriesch	oeschterich	sparrlampen
frienricht	oessterrisch	energiespaerlaempe

Figure 3: Examples of misspelled responses

verse (the number of unique tokens larger than the number of unique responses, that is, fewer recurring words than in the easiest item, LC_1). The average response length was 5 tokens.

Examples In order to illustrate the spelling errors problem, in Figure 3 we show examples of misspellings in responses to three questions which elicited simple one-word key concepts. We will use responses to these questions in one of the analyses (RAs below are reference answers provided by the teachers; vertical bar separates alternatives):

LC_1.1 Wo wohnt Alexandra?
'Where does Alexandra live?'
RA: frankreich

LC_1.6 Woher kommt Elisabeth?
'Where does Elisabeth come from?'
RA: oesterreich|wien|wien oesterreich

LC_3.1 Wie beleuchtet die Bundeskanzlerin Angela Merkel ihre Wohnung?
'How does Chancellor Angela Merkel light her apartment'
RA: energiesparlampen
'energy saving lamps'

Two of the questions (LC_1.1 and LC_1.6) appeared with the first, easiest, listening prompt. Even though identifying the answers within the audio prompts was easy for most test-takers, also low-proficiency, spelling the answers correctly turned out to be challenging, even though the elicited key concepts denote two well known European countries. The third question (LC_3.1) appeared with the last, most difficult, audio prompt and was answered by medium- to high-proficiency learners. Likewise here spelling the word is challenging. This may be partially due to the fact that "Energiesparlampe" is a compound noun.

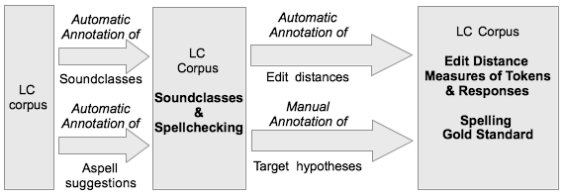

Figure 4: Corpus processing

Even this small sample illustrates the large variety of spelling errors, the high complexity of the spell-checking task, and the high demands on automated processing. Some misspellings, such as *lampel* for "lampe" or "lampen", are probably typos, while others are likely to have a phonological source, like *frankreisch* or *oesterreisch*, and among those some might be explained by interference of another foreign or the native language of the student, for instance "au" in *austereich* or "y" in *energysparen*. Some errors might be interpreted as wrong morphological forms rather than misspellings, e.g. *energisparelampen*. In many cases multiple errors are combined.

4 Spell-checking and Normalizations

As shown in Figure 4, data for analysis was prepared as follows: We created a spelling gold-standard semi-automatically by spell-checking preprocessed responses using an off-the-shelf spell-checker (described in more details in Section 4.1) and then manually annotating (verifying and correcting) the checker's outputs (Section 4.2). Each learner response and reference answer was automatically transcribed into three different phonetically-based encodings which, in the context of the automated scoring task, we treat as spelling normalizations (Section 4.3). In the analysis section we compare the spell-checked and the phonetically transcribed responses with, respectively, the strings or the transcriptions of target hypotheses and reference answers. The methods and tools used for annotation and normalization are outlined below.

4.1 Spell-checking

For automated spell-checking and spelling correction we use Aspell (Atkinson, 2006), an open source spell-checker provided by GNU. Aspell supports multiple languages and is frequently

used as a reference system in research on spell-checking and writing normalization. Crucially to this work, a large dictionary for the German language compatible with Aspell is freely available, as are implementations of the system itself. Aspell is thus a good candidate for integration into a scoring system, and so a well-motivated choice for an evaluation.

Aspell performs checking and suggests corrections based on a combination of orthographic and phonetic coding, fast dictionary lookup, and an edit distance calculation. Alternative spellings are identified by an algorithm which represents words by their orthographic forms and their "soundslike" equivalents, that is, approximate pronunciations constructed based on phonetic information. Suggestions are ordered by a weighted average of the edit distances between the candidate and the misspelled word and between the "soundslike" encodings of the two words. Aspell language versions differ in their dictionaries and phonetic data, but the underlying edit distance algorithm is the same.

Note that Aspell performs context-insensitive spell-checking, that is, individual words are processed in isolation. Thus, only non-word errors are detected, while real-word errors are not. In this study we do not address real-word errors, however, we are planning to annotate the complete data set manually in the future.

4.2 Annotation

We annotated the learner responses with target hypotheses (hypothesized intended forms) semi-automatically using the Aspell checker. For each non-word Aspell searches its dictionary and provides a list of suggested replacements. To obtain a spell-checked corpus we processed our data set with Aspell and for each word which Aspell reported as misspelled, we stored Aspell's first suggestion. Then, we manually checked the first suggestions and corrected them were necessary.

As Figure 3 illustrates, the range of spelling variants includes cases of questionable interpretation and acceptability; consider, for instance, *frienricht* or *giespallampe* as misspellings of "frankreich" and "energiesparlampe", respectively. When building the spelling gold standard we did not use the teachers' scores as guides, but rather attempted to accept generously those words which could be in good faith interpreted to be misspellings of the expected concepts. Where good-faith interpretation was impossible or borderline possible, we marked those words as uninterpretable (for instance, *frankaise*, *freikeit*, *franch* in response to LC_1.1 and *oestech*, *busterish*, *uscraisch*, or *susthei* in response to LC_1.6). We also marked foreign words explicitly (*france*, *francais*, *austria*) as some students answered in English or in their native language.

The annotation was carried out by the authors of this paper. The corpus was divided into parts and single annotation was performed for each misspelled word by one author. The manually corrected spell-checker outputs are used as a spelling gold standard. The spell-checked, annotated corpus contains 2945 responses, 15260 tokens (2898 unique responses, 2173 unique tokens).

4.3 String Normalizations

For this study we used three phonetically-based encodings: ASJP and Dolgopolsky's systems, and Soundex as baseline.

ASJPcode Automated Similarity Judgment Program (ASJP) is a procedure originating from comparative and historical linguistics developed with the view to comparing world languages by lexical similarity (Wichmann et al., 2013). Comparisons are based on word lists encoded in standardized orthography (ASJPcode), a simplified version of the International Phonetic Alphabet (International Phonetic Association, 1999). ASJP encoding consists of 41 symbols, 7 vowels and 34 consonants, which represent the commonly occurring sounds of the world's languages (for details, see Appendix C of (Brown et al., 2008)). The transcription employed in this study was specifically designed to capture the sound representations of German.

Dolgopolsky's sound classes The sound class coding system of Dolgopolsky (1986) was developed in the context of research analogous to the ASJP project, that of identifying related language families. Dolgopolsky's system groups similar consonants into 10 "sound classes" in such way that phonetic regularities within a class are more systematic than between classes. Each class is represented with a single character. Vowels are simply marked as such (V). The transcription used in this study was also designed to capture the sound system of German.

String	ASJP	Dolgopolsky	Soundex
frankeriech	fGaNkeGiS	PRVNKVRVS	F652
frankfurt	fGaNkfuGt	PRVNKPVRT	F652
fraenkerisch	fGaENkeGiS	PRVVNKVRVS	F652
fracraich	fGakGaiS	PRVKRVVS	F626
oestarreich	7oEstaGaiS	HVVSTVRVVS	O236
oestereisch	7oEsteGaiS	HVVSTVRVVS	O236
austerreich	7austEGaiS	HVVSTVRVVS	A236
austerreicht	7austEGaiSt	HVVSTVRVVST	A236

Figure 5: Examples of normalizations

Both ASJP and Dolgopolsky's transcriptions were done based on sound classes for German as is done in the LingPy package (List and Moran, 2013; List et al., 2013).

Soundex Soundex, originally patented by Russell (1918 1922), also uses sound classes to represent similar sounding words with the same encoding, however, it was designed with a practical goal of indexing family names for the census. A Soundex code represents a token with a character followed by three digits. The character denotes the first letter of the word and the digits denote the sound classes of the three following consonants. There are six such sound classes. Vowels, unless word-initial, are ignored, as are the letters H and W. If the word is longer than the four symbol sequence, the remaining letters are ignored. If it is shorter, zeros are added. Soundex is thus a more general approach than the other two and most lossy (to a greater degree abstracts away from the original string), but as it is one of the most frequently employed phonetic encodings and therefore a good baseline for comparison. Soundex has been also used in previous work on short answer scoring as a way of addressing misspellings (Hahn et al., 2013).

To illustrate the selected phonetic normalizations, examples of encoding are shown in Figure 5. As can be clearly seen, the effect of the normalizations is markedly different and reflects the more linguistically-informed basis of the ASJP and Dolgopolsky's codes: In the set of responses to LC_1.1, *frankeriech*, *fraenkerisch*, and *frankfurt* are grouped into one sound equivalence class by Soundex – an undesired result – but not by any of the other encodings. In the set of responses to LC_1.6, *oestarreich*, *oestereisch* and *austerreich*, *austerreicht* form two clusters in Soundex encoding, whereas ASJP and Dolgopolsky's codes yield more intuitive groupings; ASJP being more fine-grained than Dolgopolsky.

	Valid words	Misspelled words	Row totals
Reported	42	1040	1082
Suggestions found	21	904	925
First Correct	-	583	583
First Wrong	21	321	342
No Suggestions	21	136	157

Table 2: Performance of the Aspell spell-checker

5 Results

The following analyses are performed: We start by summarizing the performance of the spell-checker at the word-level. Next, we look at the extent of divergence of the misspelled words from the annotated target hypotheses by quantifying divergence in terms of string distances. Then, we relate misspellings and normalizations to scores: For two questions eliciting single key concept responses, we show how distance to the key concepts affects response scores. Finally, we focus on complete responses and look at relations between scores and distances between normalized learner responses and reference responses.

Two standard string distance measures are used throughout this section: Damerau-Levenshtein distance (nDL), a variant of Levenshtein edit distance which accounts for transposition of adjacent characters (Damerau, 1964; Levenshtein, 1966), and string vector cosine based on n-grams. A length correction on the edit distance is performed in a standard way by dividing the distance by the length of the longer string. Cosine similarity is computed for unigrams, bigrams and trigrams. Because the data is not normally distributed and for some items the number of observations is low, instead of performing statistical analysis, we present boxplots to show general tendencies in an informative way.

5.1 Automated Spell-checking

The performance of the Aspell spell-checker against the gold-standard is summarized in Table 2. "Valid words" refers to correctly spelled words and "Misspelled words" to non-words. The numbers refer to *unique* tokens.

Out of the 2173 unique tokens, Aspell reported around 50% (1082) as misspelled. Since there were 1818 occurrences of misspellings overall, it is clear that a lot of the same misspellings recur. Out of the 1082 reported misspellings

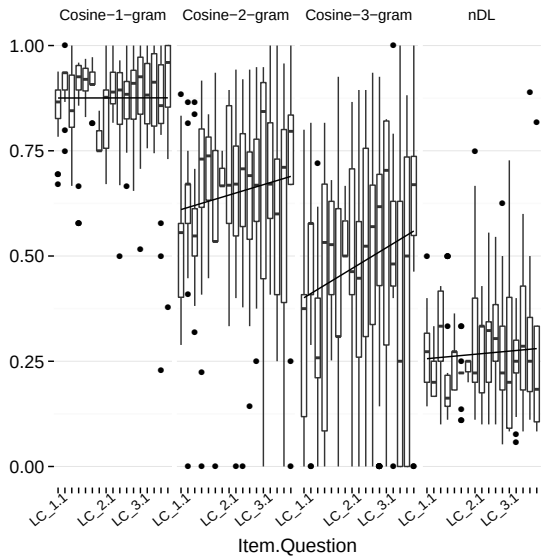

Figure 6: Per item distribution of distances between misspelled words and target hypotheses

Aspell reported 21 (4%) correctly spelled words as misspelled and suggested a correction (false positives). Overall Aspell's precision in identifying misspellings in our data is thus at 96%.[1]

Now, as far as automated correction is concerned, suggestions were found for not even 60% of the tokens. Out of the 925 tokens for which suggestions were found, 321 first suggestions were wrong, yielding a false negative rate of 64%. With 321 wrong suggestions and 136 cases for which suggestions were not available, about 45% of the non-word misspellings are not accounted for correctly by Aspell. These results are similar to those reported by Rimrott and Heift (2008).

A major issue for Aspell, and, as can be expected, for any off-the-shelf German spell-checker, are compound nouns. Two of the listening prompts contained compounds as key concepts: "Marxhaus" in the answer to *Where are Peter and Birgit?* (RA: 'In front of Marx' birth place in Trier') and "Energiesparlampen" in the answer to the previously mentioned LC_3.1. "Marxhaus" is not in Aspell's dictionary; the closest suggestions it finds as replacements include *Matthäus* (Matthew; as in Matthew the Apostle), *Parkhaus* (carpark) or even *Hausbar* (house bar). Compounds account for all the 21 valid words which Aspell identified as misspellings.

[1]We cannot provide recall results at this point since our gold standard includes only non-words identified by Aspell. We are planning to annotate real-word errors in the future.

Most of the remaining errors are due to context insensitivity; for instance, to "What did Karl Marx do in Cologne?" (RA: "Leitung der Neuen Rheinischen Zeitung" ('Led the "New Rhinish Newspaper"') a student wrote: *radikal demochratisch behzatung* ('radical democratic UNINTERPRETABLE') for which Aspell suggested *radikal demokratische Beratung* (radical democratic counseling) which considering pure edit distance obviously makes sense, otherwise not.

5.2 Diversity of Misspellings

Figure 6 shows the distribution of cosine and normalized Damerau-Levenshtein distances (nDL) to target hypotheses with linear trend lines. On the x-axis, items within distance measure groups are ordered as in Figures 1 and 2. As can be seen in the plots, the range of unigram cosine values is large for some items. Thus a lot of misspellings involve more than just letter transpositions. The large ranges in bigram cosines and many values at 0 for trigrams show that misspellings tend to diverge from the target hypotheses to a large extent.

For the easier questions (left end of the x subaxes) the ranges of unigram cosine and Levenshtein distance tend to be smaller, while bigram and trigram cosines are larger and they are also closer to the low-end of the scale. This means that in the easy questions, misspellings tend to contain the right letters, but the letters are misplaced. The same can be seen for the difficult questions (except for the last one). The intermediate difficulty items tend to have the least letter overlap and many trigram similarities at the low end of the scale. These are likely to be most difficult to correct automatically, but possibly easier to identify as qualifying to be scored at 0.

5.3 Relation to Scores: Misspelled Key Concepts

As mentioned in Section 3, we used responses to two questions which elicited one key concept, LC_1.1 and LC_1.6, to investigate the relation between misspellings and scores. From the LC_1.1-LC_1.6 corpus subset, we extracted responses which contained tokens with gold standard annotation corresponding to the expected concept: "frankreich" for LC_1.1 and "oesterreich" for LC_1.6. There were 236 and 260 such responses, respectively.

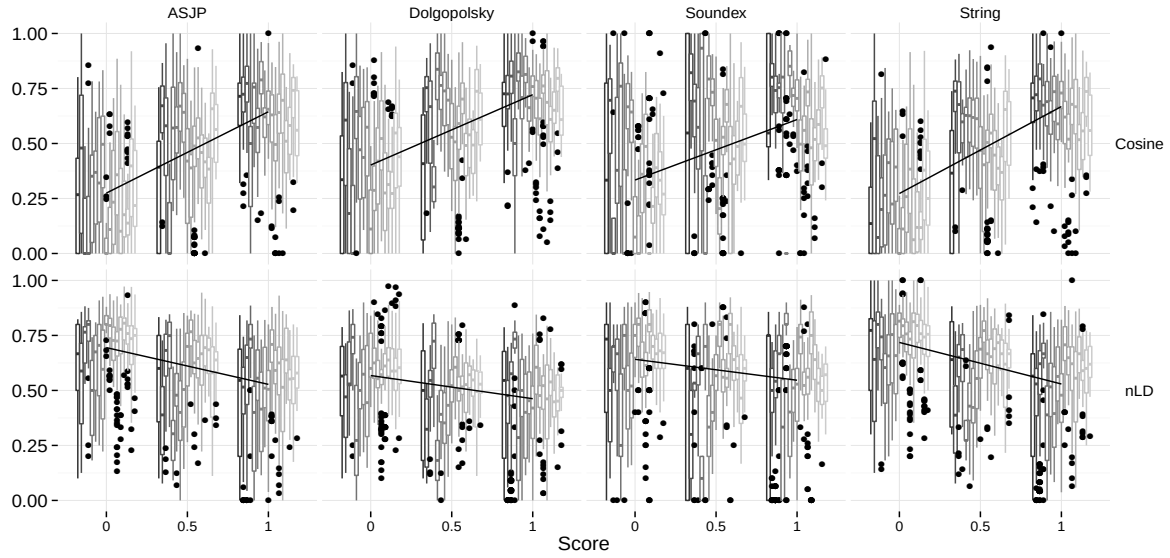

Figure 8: Per score distribution of distances between normalized responses and reference responses

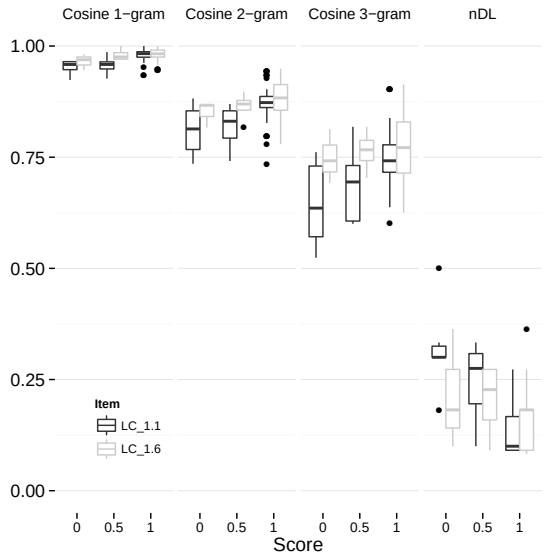

Figure 7: Per score distribution of distances between misspelled key concepts and target hypotheses for two items

For these responses, in Figure 7 we show the distribution of the distances to the target hypotheses between score points.

Most of the expected general tendencies can be found in the data: cosine distances for all n-grams increase with the scores as expected. Levenshtein distance decreases as expected for LC_1.1, but the pattern for LC_1.6 is not clear. Moreover, and more interestingly, the acceptability thresholds for the two questions appear to be different. Responses with misspelled key tokens of lower similarity to the target concept tokens are accepted with partial and full scores in LC_1.1. Also a larger range of similarity accounts for partial and full points in LC_1.1. This suggests that what counts as acceptable in terms of misspellings could be item-specific and different thresholds would have to be used for different items.

5.4 Relation to Scores: Normalizations

Finally, we investigate the relation between sound class-based response normalizations and the scores assigned by teachers. Complete preprocessed learner and reference responses have been transcribed into the three encodings described in Section 4.3. Based on Figure 7 the 3-gram cosine distance yields a pattern that best distinguishes between the three score points. Therefore, only 3-gram cosine distances are reported for the normalized responses. We seek to find out which normalization yields the most consistent patterns in terms of the expected relation to the teachers' scores.

The distributions of distances between normalized learner and reference responses for all the items are shown in Figure 8. Items clustered by score-point are ordered as in Figures 1 and 2. Distribution of string distances is shown for comparison. Linear trends are overlayed.

Two immediate observations can be made of the results. First, the score-based grouping is not clear-cut and the distance ranges overlap across score levels. Second, the expected pattern of cosine distance (linearly) increasing and normalized

Levenshtein distance (linearly) decreasing can be seen in the distribution of ASJP and Dolgopolsky normalizations, but less so in the distribution of Soundex distances across items. Soundex transcriptions do not distinguish well between the scores based on Levenshtein distance and only somewhat better based on cosine; for most items there is little difference between mean distances for scores 0.5 and 1 on the nLD measure and between mean scores 0.5 and 1. ASJP and Dolgopolsky normalizations are more stable in terms of variance, with ASJP, moreover, displaying fewer outliers. This confirms our hypothesis that the more linguistically-informed encoding yields clusters which better correspond to the assigned scores. It also suggests that these encodings might result in better performance on the automated scoring task. We are planning to investigate this in the course of further work. The ASJP and Dolgopolsky distributions moreover better reflect the pattern of string-based distances than the Soundex distributions. Finally, ASJP and Dolgopolsky normalizations appear more stable across items on both distance measures and the shape of the distributions is similar. It is possibly a combination of both that would work best as features for scoring.

6 Conclusions and Further Work

We presented a study on misspellings in a corpus of constructed responses to listening comprehension items used for placement testing for German. Not surprisingly, our data contains a large number of misspellings (around 50% of the unique words that learners used). The first-ranked suggestions of an off-the-shelf spell-checker were correct in not even 60% of the cases. This is likely to be partially due to the fact that the range of divergence from target forms is substantial. It also varies between questions. The majority of false positives were due to compounds specific to the listening prompts. An obvious solution we are pursuing to improve precision and reduce false negative suggestion rate is constructing two dictionaries: one prompt-specific and the other learner-language specific; the purpose of the latter is to provide prompt-specific frequent invalid forms produced by the learners.

We have also shown that while in general the expected trend in scoring misspelled responses can be observed, however, acceptability of divergence from target forms appears to be item-specific. Finally, we proposed sound class-based normalizations as a method of grouping noisy responses in terms of their pronunciation similarity as well as related distances between normalized responses and reference answers to response scores. This served to evaluate prospects for a normalization-based approach to response clustering. Soundex, the most frequently employed normalization, does not distinguish between responses at different score-points, so it can be considered the worst choice for a normalization-based approach. Both of the more elaborate phonetic transcriptions, based on ASJP's and Dolgopolsky's codes, perform better than Soundex and are promising directions to pursue. We will experiment with including distances to reference answers based on both representations as features for (semi-)automated scoring.

Acknowledgments

We thank Dr. Kristin Stezano Cotelo from the Saarland University International Office for collaboration on placement testing thanks to which this research is possible. We would like to thank Johannes Dellert for letting us use his code for sound class-based normalizations. We also thank the three anonymous reviewers for their helpful comments.

This work was funded by the Ministry of Science, Research and the Arts of Baden-Württemberg within the FRESCO project. Magdalena Wolska is supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63).

References

Kevin Atkinson. 2006. Gnu Aspell 0.60.7. `http://aspell.net`.

Adriane Boyd. 2010. EAGLE: an Error-Annotated Corpus of Beginning Learner German. In *Proceedings of the 7th LREC*. `http://www.lrec-conf.org/proceedings/lrec2010/summaries/812`.

Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(4):285–308. doi:10.1524/stuf.2008.0026.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*. doi:10.1145/363958.363994.

Aharon B. Dolgopolsky. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia. In *Typology, Relationship and Time: A Collection of Papers on Language Change and Relationship by Soviet Linguists*, pages 27–50. (Original: 1964 In: *Voprosy Jazykoznanija* 2).

Michael Flor and Yoko Futagi. 2012. On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP*. http://dl.acm.org/citation.cfm?id=2390397.

Michael Hahn, Niels Ott, Ramon Ziai, and Detmar Meurers. 2013. CoMeT: Integrating different levels of linguistic modeling for meaning assessment. https://aclweb.org/anthology/S/S13/S13-2102.pdf.

Trude Heift and Anne Rimrott. 2008. Learner responses to corrective feedback for spelling errors in CALL. *System*. doi:10.1016/j.system.2007.09.007.

International Phonetic Association, editor. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Vilius Juozulynas. 2013. Errors in the compositions of second-year german students: An empirical study for parser-based icali. *CALICO Journal*. https://ns.calico.org/html/article_578.pdf.

V.I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10(8):707–710.

Johann-Mattis List and Steven Moran. 2013. An open source toolkit for quantitative historical linguistics. In *Proceedings of the ACL Conference (System Demonstrations)*. https://www.aclweb.org/anthology/P/P13/P13-4003.pdf.

Johann-Mattis List, Steven Moran, Peter Bouda, and Johannes Dellert. 2013. LingPy. Python Library for Automatic Tasks in Historical Linguistics. http://www.lingpy.org.

Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing target hypotheses in the Falko corpus. In *Automatic Treatment and Analysis of Learner Corpus Data*, volume 59 of *Studies in Corpus Linguistics*, pages 101–123.

Anne Rimrott and Trude Heift. 2008. Evaluating automatic detection of misspellings in German. *Language Learning & Technology*, 12(3):73–92. http://llt.msu.edu/vol12num3/rimrottheift.pdf.

Robert C. Russell. 1918, 1922. US Patents No.: 1261167 and 1435663. (Retrieved 04/15 via http://patft.uspto.gov/netahtml/PTO/search-adv.htm).

Howida A. Shedeed. 2011. A new intelligent methodology for computer based assessment of short answer question based on a new enhanced soundex phonetic algorithm for arabic language. *International Journal of Computer Applications*. http://research.ijcaonline.org/volume34/number10/pxc3876054.pdf.

Søren Wichmann, André Mller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Zarina Molochieva, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitry Egorov, Matthias Urban, Robert Mailhammer, Agustina Carrizo, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant, and Pilar Valenzuela. 2013. The ASJP-Database (version 16). http://asjp.clld.org (Retrieved 04/15).

Magdalena Wolska, Andrea Horbach, and Alexis Palmer. 2014. Computer-assisted scoring of short responses: the efficiency of a clustering-based approach in a real-life task. In *Advances in Natural Language Processing (Proceedings of the 9th International Conference on Natural Language Processing (PolTAL-14))*. doi:10.1007/978-3-319-10888-9_31.