A Language Resource of German Errors Written by Children with Dyslexia

Maria Rauschenberger [1], **Luz Rello** [2], **Silke Füchsel**[3], **Jörg Thomaschewski**[3]
[1] Web Research Group, Universitat Pompeu Fabra
[2]Human-Computer Interaction Institute, Carnegie Mellon University
[3]University of Applied Science Emden/Leer
Maria.Rauschenberger@upf.edu, luzrello@cs.cmu.edu, sfuechsel@gmx.de, joerg.thomaschewski@hs-emden-leer.de

Abstract

In this paper we present a language resource for German, composed of a list of 1,021 unique errors extracted from a collection of texts written by people with dyslexia. The errors were annotated with a set of linguistic characteristics as well as visual and phonetic features. We present the compilation and the annotation criteria for the different types of dyslexic errors. This language resource has many potential uses since errors written by people with dyslexia reflect their difficulties. For instance, it has already been used to design language exercises to treat dyslexia in German. To the best of our knowledge, this is first resource of this kind in German.

Keywords: Written Errors, Errors, Dyslexia, Visual, Phonetics, Resource, German

1. Introduction

Dyslexia is a specific learning disability with neurological origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the perception of visual and auditory components of language.

General misspells have already proven to be a useful source of knowledge for various applications (Gelman and Barletta, 2008; Piskorski et al., 2008; Baeza-Yates and Rello, 2012). A list of annotated errors of children with dyslexia in German is a useful resource because the errors that people with dyslexia make reflect the types of difficulties that they have (Sterling et al., 1998). As a matter of fact, these type of written errors have been used for various purposes such as studying dyslexia (Aragón and Silva, 2000; Connelly et al., 2006), diagnosing dyslexia (Schulte-Körne et al., 1996; Toro and Cervera, 1984), for build tools to treat[1] and to create applications to support dyslexia, such as dyslexia screeners (Rello et al., 2016b), spellcheckers (Korhonen, 2008; Pedler, 2007; Rello et al., 2015), text prediction software[2] or spelling exercises (Rauschenberger et al., 2015; Rello et al., 2014b). There are similar errors resources for English (Pedler, 2007) and Spanish (Rello et al., 2014a; Rello et al., 2016a) but, to the best of our knowledge, this is first resource of this kind in German.

There are similar errors resources for English (Pedler, 2007) and Spanish (Rello et al., 2014a; Rello et al., 2016a) but, to the best of our knowledge, this is first resource of this kind in German.

In this paper, we present the creation of a new resource composed of German errors written by people with dyslexia that did not exist before. This involved the collection and the annotation of the errors with different kind of information, such as, phonological and visual information; and the creation of new categories specifically for German language. The annotation criteria had to be adapted for German because it is a language with a different orthography and syllabic structure (Seymour et al., 2003). The resource of dyslexic errors are available on-line.[3]

2. Collecting Errors

We collected 47 texts (homework exercises, dictations, and school essays) written by students from 8 to 17 years old. In Figure 1 we show an example of a handwritten text from a 10 year-old boy with dyslexia. We kept collecting texts until we reached 1,000 written errors by people with dyslexia. Previous research have shown that around thousand errors are enough to extract for useful conclusions (Pedler, 2007; Rauschenberger et al., 2015; Rello et al., 2014b).

A total of 32 texts came from children who have been diagnosed with dyslexia. The remaining 15 texts came from students with a high spelling error rate that were chosen by their teachers. The students attended either primary school, comprehensive school (*Gesamtschule*), high school (*Gymnasium*) or a school for children with learning difficulties (*Förderschule*).

3. Error Classification

We analyzed the errors and define two more error categories specific to German: capital letter and non-capital letter errors. The rest of the errors were consistent with Pedler's classification of dyslexic errors (Pedler, 2007).[4] The error categories are the following:

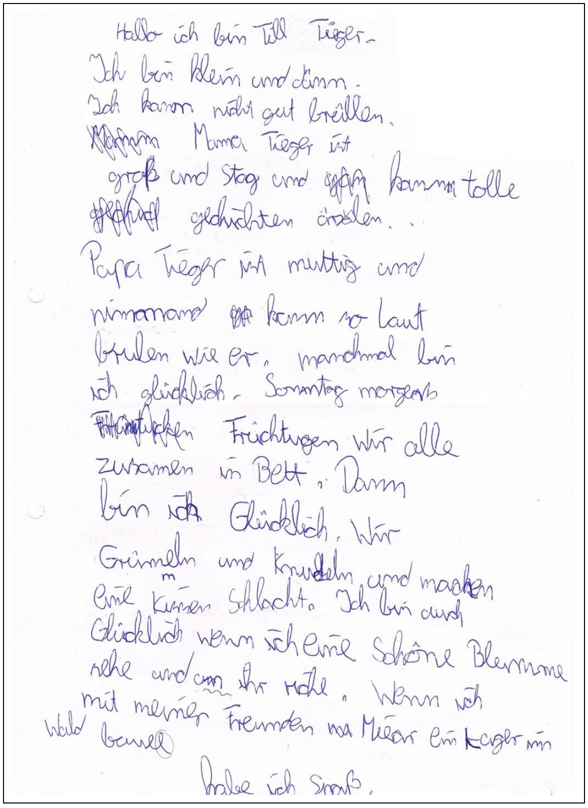
- **Substitution.** Changing one letter for another, for example **grümeln (krümeln, 'crumble').*
- **Insertion.** An insertion of one letter, such as **muttig (mutig, 'bravely').*
- **Omission.** An omission one letter, as in **zusamen (zusammen, 'together').*

[1] *Dyseggxia* is available at `https://itunes.apple.com/de/app/dyseggxia/id534986729?mt=8`.

[2] *Penfriend XL* is available at `http://www.penfriend.biz/`.

[3] The resource is available at `http://goo.gl/LRaUDA` .

[4] Examples with errors are preceded by an asterisk '*'. We use the standard linguistic conventions: '<>' for graphemes, '/ /' for phonemes and '[ ]' for phones.



Hallo ich bin Till Tieger.
Ich bin klein und dünn.
Ich kann nicht gut brüllen.
Mama Tieger ist
groß und Stag und kannn tolle
gechichten ärzälen.
Papa Tieger ist muttig und
nimmand kann so laut
brulen wie er. manchmal bin
ich glücklich. Sonntag morgens
Früchtugen wir alle
zusamen in Bett. Dann
bin ich Glücklich. Wir
Grümeln und Knudeln und machen
eine Kissen Schlacht. Ich bin auch
Glücklich wenn ich eine Schöne Blumme
sehe und an ihr riche. Wenn ich
mit meiner Freunden Miear ein Lager im
Wald bauee
habe ich Spaß.

Hallo ich bin Till *Tieger.
Hallo ich bin Till Tiger.
'Hi, I am Till Tiger.'

Ich bin klein und dünn. Ich *kannn nicht gut brüllen.
Ich bin klein und dünn. Ich kann nicht gut brüllen.
'I am small and thin. I cannot roar good.'

Mama *Tieger ist groß und *Stag und kannn tolle *gechichten ärzälen.
Mama Tiger ist groß und stark und kann tolle Geschichten erzählen.
'Tiger Mum is tall and strong and can tell fantastic stories.'

Papa *Tieger ist *muttig und *nimmand kann so laut *brulen wie er.
Papa Tiger ist mutig und niemand kann so laut brüllen wie er.
'Tiger Dad is brave and nobody can roar as load as he.'

***manchmal bin ich glücklich.**
Manchmal bin ich glücklich.
'Sometimes I am happy.'

***Sonntag morgens *Früchtugen wir alle *zusamen *in Bett.**
Sonntag morgens Frühstücken wir alle zusammen im Bett.
'Sundays, we all have breakfast in the bed.'

Dann bin *ick *Glücklich. Wir Grümeln und Knudeln und machen eine *Kissen *Schlacht.
Dann bin ich glücklich. Wir krümeln und knuddeln und machen eine Kissenschlacht.
'Then I am happy. We crumble and cuddle and make apillow fight'

Ich bin auch *Glücklich wenn ich eine *Schöne *Blumme sehe und an ihr *riche.
Ich bin auch glücklich wenn ich eine schöne Blume sehe und an ihr rieche.
'I am happy too, if I see a beautiful flower and I smell her.'

Wenn ich mit meiner *Freunden *Miear ein *lager im Wald *bauee habe ich Spaß.
Wenn ich mit meiner Freundin Mia ein Lager im Wald baue, habe ich Spaß.
'I have fun, when I am making a camp in the woods with Mia.'

Figure 1: Example of a handwritten text of a 10 year-old boy with dyslexia (left) and its transcription in German and English (right).

- **Transposition.** Reversing the order of two letters, for example *Porblem (Problem, 'problem').*
- **Multi-errors.** They differ in more than one letter from the target word such as **Stag (stark, 'strong').*
- **Word boundary errors.** They are run-ons and split words. A run-on is the result of omitting a space, such as *nichtärgern (nicht ärgern, 'don't tease').* A split word occurs when a space is inserted in the middle of a word, such as *Vogel futter (Vogelfutter, 'bird food').*
- **Capital Letter.** In German nouns are written with capital letters, while other kinds of words like verbs, adjectives or articles are not (Stang, 2010). For example **geschichten (Geschichten, 'stories').*
- **Wrong Capital Letter.** As explained before, it is confusing for children to decide which word has to be written with or without capital letters. For instance, verbs, adjectives or articles are not written with capital letters as in **Glücklich (glücklich, 'happy').*

4. Error Annotation

We annotated each of the word-error pairs with linguistic features and created new categories for German. Each of the word-error pairs was enriched with meta data and was classified as the following:

- **Unique numbering**: unified number to distinctly identify the data.
- **Target word**: word the person aimed to write.
- **Misspelled word**: the wrongly written word.
- **Damerau-Levenshtein distance**: the minimum number of edits (insertion, deletion, substitution, transposition) required to change the misspelled error into the (target) correct word (Damerau, 1964; Levenshtein, 1965).[5]
- **Target and misspelled word frequencies**: defined as the number of hit counts in a major search engine [6] for the frequencies of the target and misspelled word. The search engine does not distinguished between non-capital and capital letters. Therefore words which only

[5] The Levenshtein distance (Levenshtein, 1965) is the minimum number of substitutions, insertions and deletions to transform one string into another. The Damerau version (Damerau, 1964) counts a transposition as a single error instead of two errors.

[6] Here we refer to all web pages written in German and not only web pages from Germany. For determining whether a web page was written in German, we used Google Advanced Search settings (http://www. google.com/advanced_search).

differ through a capital letter the same frequency *e.g.* **hubschrauber* (*Hubschrauber*, *'helicopter'*) have.

- **Target and misspelled length**: number of characters the target word and the error word have.
- **Error position**: the position in the target word where the error occurs.
- **Syllable error**: the position of the syllable in the target word where the error occurs.
- **Target word syllables**: number of syllables.[7]
- **Target syllable**: the structure of the syllable where the error occurs, such as C(onsunant)V(owel), CVC, or CCV, among others.
- **Type of error**: The errors were tagged according to the classification presented in Section 3.
- **Real word**: this Boolean attribute records if the error produced another real word. For example *Schal ('scarf')* and *Schall ('sound')*.
- **First letter error**: this Boolean attribute records if the error is produced in the first letter of the word, for instances *föllig (völlig, 'fully')*.
- **Last letter error**: this Boolean attribute records if the error is produced in the last letter of the word such in *dan (dann, 'then')*.
- **Correct Letter** and **Error Letter**: The correct letter is the letter that was mistaken in the correct word by the **Error Letter**.

4.1. Visual Features

For each target and error grapheme we annotate the letters involved in the error with the following visual information, considering handwritten text (Table 1).
Four handwriting alphabets are commonly used in German schools (Topsch, 2005; Bartnitzky, 2010). These are the *Lateinische Ausgangsschrift, Vereinfachte Ausgangsschrift, Schulausgangsschrift* and *Grundschrift*. In some states there is one mandatory alphabet to be used by the school, while in other states schools can decide. For our method we choose the *Lateinische Ausgangsschrift* (Topsch, 2005), shown in Figure 2, because it is commonly used in schools where the texts were collected.

- **Mirror letter**: Boolean attribute that indicates if the mirror of a letter produces another letter, such as <d> and <b> or <m>, and <w>.
- **Rotation**: Boolean attribute that indicates if the rotation of a letter produces another letter, such as <d> and <p>.
- **Fuzzy letters**: Boolean attribute that indicates if the letter has similar visual letters (not due to rotate or mirror) such as <s> and <z>.

Feature	Letters
Mirror	**Yes** = <b, p, d, q, m, w, u, n, v, H> **No** = rest of letters
Rotation	**Yes** = <b, g, h, y, p, d, H> **No** = rest of letters
Fuzzy	**Yes** = <a, o, ä, ö, b, d, g, h, m, n, p, q, u, v, w, y>

Table 1: Visual features of the annotated target and error letters.

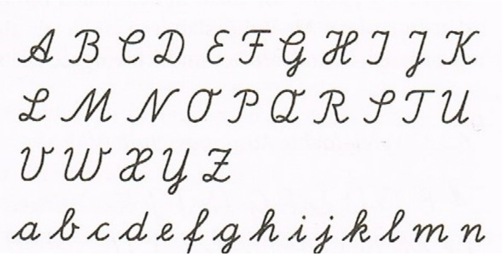


Figure 2: *Lateinische Ausgangsschrift*, a handwriting alphabet commonly used in German schools.

4.2. Phonetic Features

Each of the error words were tagged using a scale inspired by the error analysis of the DRT (Grund et al., 2004). This scale is based on traditional articulatory phonetic features (International Phonetic Association, 1999) and is divided into the following categories.

- **Sound distinction.** This category has two values: similar sound errors, *e.g.* **eingebackt [‘aɪngebakt]* (*eingepackt [‘aɪngepakt]*, *'wrapped'*) and different sound errors, *e.g.* **Tüsch* [tyʃ] (*Tisch* [tɪʃ], *'table'*).
- **Sound sequence.** The category has three values: error words with missing phonemes, *e.g.* **Mächen* (*Märchen*, *'fairy tale'*); added phonemes, *e.g.* **Spieln* (*Spiel*, *'game'*) or transposition of letters, *e.g.* **Porblem* (*Problem*, *'problem'*).
- **Combination of consonants.** Some consonants are pronounced in a different way when they are combined with each other. For example the consonant <s> [s] and <p> [p] are pronounced like *[ʃp]* when they are written together.
- **Words with** **<v>.** Words written with a <v> since its sound correspondence is not transparent in German, *e.g.* **Ferkäuferin (Verkäuferin, 'seller')*.
- **Umlaut.** There are three umlauts in the German language <ä; ü; ö>. The dots are often missing in texts.
- **Double consonant / false double consonant.** After a short, stressed vowel there are usually two or more consonants following. If there is only one consonant following, this one should be doubled most of the times, *e.g.* **vergesen* (*vergessen, 'forget'*). Double consonants also appear at syllable boundaries. This

[7] The syllables where checked with http://www.duden.de.

category include false double consonants and double consonants in the wrong place, such as **Unffall* (*Unfall*, *'accident'*).

- **Lengthening.** There are different types of lengthening for a vowel in German. This process gives as a result a long stressed vowel. The long vowel <i> [i:] is frequently lengthened with an *e* which is not audible. A typical error is **wider* (*wieder [vi:dɐ], 'again'*). About 20% of the long, stressed vowels are lengthened with a <h>, *e.g.* **erzälen* (*erzählen [ɛr'tsɛ:lən], 'tell'*) (Grund et al., 2004). A few long stressed vowels are lengthened by double vowels like **Hare* (*Haare ['ha:rə], 'hair'*). This category include false lengthening errors produced by adding <e> or <h> after a long stressed vowel in the wrong place, *e.g.* **währe* (*wäre ['vɛ:rə], 'would be'*).

- **Derivation.** Related words that are often written the same way or similar but pronounced different. To write these words in the right way, one possibility is to have a look at the plural form so that the right writing can be derived, *e.g.* *Walt* (*Wald [valt]; Wälder [vɛldɐ], 'forest'*).

  Words with <s/β>. A word with a voiced [s] is always written with an <s>. Words with a voiceless [s] have specific rules which determine if they have to be written with <s> <ss> or <β>, *e.g.* **Reisverschlus (Reiβverschluss, 'zipper')*.

5. Conclusions

In this paper we have presented the compilation and the annotation criteria of a list of 1,021 unique errors written by people with dyslexia in German. The adaptation of a Spanish based method to the German language raised a number of challenges. For instance, the handwriting systems taught in schools in Germany are different from the Spanish ones, so the visual features needed to be redefined. We annotated each of the word-error pairs with linguistic features and two new error categories were specially created for the German language. We are planning to use the resource for the detection of dyslexia in German (Rauschenberger, 2016) using web applications.

Acknowledgements

We deeply thank Janka Melgert-Retelsdorf and all children for making the collection of textes possible. We also thank Hendrik Witzel for helping with the phonetic error annotation.

Bibliographical References

Aragón, L. E. and Silva, A. (2000). Análisis cualitativo de un instrumento para detectar errores de tipo disléxico (IDETID-LEA) (Qualitative analysis of an instrument to detect dyslexic errors, IDETID-LEA). *Psicothema*, 12(Supl. 2):35–38.

Baeza-Yates, R. and Rello, L. (2012). On measuring the lexical quality of the web. In *The 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 1–6, Lyon, France.

Bartnitzky, H. (2010). Grundschrift - damit kinder besser schreiben lernen. In Grundschulverband, editor, *Grundschulverband aktuell*, volume 110, pages 4–8. Grundschulverband.

Connelly, V., Campbell, S., MacLean, M., and Barnes, J. (2006). Contribution of lower order skills to the written composition of college students with and without dyslexia. *Developmental Neuropsychology*, 29(1):175–196.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the A.C.M.*, 7:171–176.

Gelman, I. A. and Barletta, A. L. (2008). A "quick and dirty" website data quality indicator. In *The 2nd ACM Workshop on Information Credibility on the Web (WICOW '08)*, pages 43–46, Napa Valley, USA.

Grund, M., Naumann, C. L., and Haug, G. (2004). *Diagnostischer Rechtschreibtest für 5. Klassen: DRT 5 ; Manual*. Deutsche Schultests. Beltz Test, Göttingen, 2., aktual. aufl. in neuer rechtschreibung edition.

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

Korhonen, T. (2008). Adaptive spell checker for dyslexic writers. In *Proceedings of the 11th international conference on Computers Helping People with Special Needs (ICCHP '08)*, pages 733–741, Berlin, Heidelberg. Springer.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

Pedler, J. (2007). *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Ph.D. thesis, Birkbeck College, London University.

Piskorski, J., Sydow, M., and Weiss, D. (2008). Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '08)*, pages 25–28, New York, NY. ACM Press.

Rauschenberger, M., Füchsel, S., Rello, L., Bayarri, C., and Thomaschewski, J. (2015). Exercises for German-Speaking Children with Dyslexia. In *Human-Computer Interaction–INTERACT 2015*, pages 445–452. Springer International Publishing.

Rauschenberger, M. (2016). Dysmusic: Detecting dyslexia by web-based games with music elements. In *Proc. Web4All'16*, Montreal, Canada. ACM Press.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2014a). DysList: An annotated resource of dyslexic errors. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1289–1296, Reykjavik, Iceland, May.

Rello, L., Bayarri, C., Otal, Y., and Pielot, P. (2014b). A computer-based method to improve the spelling of children with dyslexia using errors. In *Proc. The 16th International ACM SIGACCESS Conference of Computers and Accessibility (ASSETS 2014)*, Rochester, USA, October.

Rello, L., Ballesteros, M., and Bigham, J. (2015). A spellchecker for dyslexia. In *Proc. ASSETS'15*, Lisbon, Portugal. ACM Press.

Rello, L., Baeza-Yates, R., and Llisterri, J. (2016a). A resource of errors written in spanish by people with dyslexia and its linguistic, phonetic and visual analysis. *Language Resources and Evaluation*.

Rello, L., Ballesteros, M., Ali, A., Serra, M., Alarcón, D., and Bigham, J. P. (2016b). Dytective: Diagnosing risk of dyslexia with a game. In *Proc. Pervasive Health'16*, Cancun, Mexico.

Schulte-Körne, G., Deimel, W., Müller, K., Gutenbrunner, C., and Remschmidt, H. (1996). Familial aggregation of spelling disability. *Journal of Child Psychology and Psychiatry*, 37(7):817–822.

Seymour, P. H. K., Aro, M., and Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94(2):143–174.

Stang, C. (2010). *Duden, Deutsche Rechtschreibung*. Praxis kompakt. Dudenverl, Mannheim and Leipzig and Wien and Zürich.

Sterling, C., Farmer, M., Riddick, B., Morgan, S., and Matthews, C. (1998). Adult dyslexic writing. *Dyslexia*, 4(1):1–15.

Topsch, W. (2005). *Grundkompetenz Schriftspracherwerb: Methoden und handlungsorientierte Praxisanregungen*, volume Bd. 5 of *Beltz Pädagogik*. Beltz, Weinheim and Basel, 2., überarb. und erw. aufl edition.

Toro, J. and Cervera, M. (1984). *TALE: Test de Análisis de Lectoescritura (TALE: Literacy Analysis Test)*. Visor, Madrid.