Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language xxx (2017) xxx-xxx

www.elsevier.com/locate/csl

Automatic orthographic error tagging and classification for German texts☆

Q1 Kay Berkling[a,*], Rémi Lavalley[b]

Q2 [a] Cooperative State University, Kalrsruhe, Germany
[b] Inline Internet GmbH, Karlsruhe, Germany

Received 30 April 2016; received in revised form 18 July 2017; accepted 8 November 2017
Available online xxx

Abstract

This paper evaluates an automatic spelling error tagger and classifier for German texts. After explaining the existing error tags in detail, the accuracy of the tool is validated against a publicly available database containing around 1700 written texts ranging from first grade to eighth grade. The tool is then applied to a longitudinal study consisting of weekly children's texts from second and third grades. It can be shown which error categories contribute most significantly to children's error profiles. Additionally, it can be shown whether or not children make progress on improving in the categories under study.

Keywords: Orthography; Children's text; Corpus; Longitudinal study

1. Overview

This paper motivates, describes and evaluates an automatic spelling error classification system for German writing of unconstrained texts. A database with 1700 spontaneously written texts from grades 1–8, including *Grundschule, Hauptschule* and *Realschule* is used for evaluation of the algorithm. The annotation and study of spelling errors represents one dimension towards gaining a deeper understanding about children's writing acquisition. For this purpose, a second corpus from a longitudinal study allows us to looks at the orthographic development of children's writing in second and third grade over a period of three months by applying the tool.

The rest of the paper is structured as follows. The introduction will give an overview of the field of orthographic acquisition in Section 2 followed by a theoretical introduction to orthographic issues that arise in the German spelling system according to their phonemic, syllabic, morphological and sentence related derivation in Section 3. In Section 4 we then proceed to describe the system that will be used to prepare the data for the classification. The algorithms for tagging and classifying the spelling errors are described in Section 5. After a brief review of the corpora in Sections 6 and 7 evaluates the quality of the automatic spelling error annotation of the data. Section 8 presents results on the longitudinal data using the evaluated tool. Section 9 concludes this paper and talks about future work.

☆ This paper has been recommended for acceptance by Roger Moore.
* Corresponding author.
E-mail address: berkling@dhbw-karlsruhe.de (K. Berkling).

http://dx.doi.org/10.1016/j.csl.2017.11.002
0885-2308/ 2017 Published by Elsevier Ltd.

Please cite this article as: K. Berkling, R. Lavalley, Automatic orthographic error tagging and classification for German texts, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.11.002

2. Introduction and related work

Reading and spelling are key skills acquired by children during their first four years of school. These two skills are tightly interlaced. By studying one skill we deepen our knowledge into both skill acquisitions (Retelsdorf and Köller, 2014). According to PISA, IQB (Stanat et al., 2016) and IGLU (OECD, 2014), a significant number of school children are still left behind in Germany. PISA (2000−2012) has documented a significant discrepancy between students' scores. It is generally known that underachievement in reading and spelling acquisition can stem from a lack of a variety of skills, including phonemic awareness, knowledge of grapheme-phoneme correspondences and reading (Read, 1975; Bissex, 1985; Wagner et al., 1994; Treiman, 1993). In addition to that, Germany has a number of children with migrant backgrounds, only those with foreign nationality entering the official statistics. Despite almost two decades of effort to increase the number of foreign children participating at the level of *Gymnasium* (high school), the statistics since 1995 have not changed much.

In order to prevent problematic developments, early reading, spelling and language skills have to be targeted in specific interventions (National Reading Panel et al., 2000). Especially reading and spelling interventions administered from Grade 1 to Grade 2 show positive effects (Suggate, 2014). In order for instructional material, diagnostics and intervention to be effective, more research is needed to understand how writing skill acquisition develops.

Analyzing orthographic abilities of children in Germany are usually either performed on smaller datasets (Berkling et al., 2011), or the spelling errors are evaluated at a high level (Thomé, 1999) and are marked by hand (Hanke and Schwippert, 2005), or orthographic progress is marked in broad steps of acquisition (Sassenroth, 2000; Bredel, 2011). Studies often focus on children with dyslexia or multilingualism (Günther et al., 1989; Landerl, 1996; Röber-Siekmeyer, 2003). A large body of research has mostly focused on phonological awareness and its effect on spelling capability (Roth and Schneider, 2002; Küspert, 1998; Reichardt, 2015). Using hand-labels, error categories tend to be at a broader level (miss-spellings at word level, similar to a spell-checker, or counting of missing capitalization as examples) or necessitate an incredible amount of work and estimations for normalization purposes in order to cover detailed analysis. Unlike marking a mistake like a spell checker, such detail relates to the underlying linguistic conception (a phoneme-level mistake vs. morpheme-level derivational mistake "Hant" or "Hende" as opposed to "Hand" and "Hände").

In order to capture performance on a detailed list of error categories, a number of pencil and paper tests have been developed as standardized tests with large data collections to form statistically accurate diagnoses, normed for specific grade levels. Among these are the "Diagnostische Rechtschreibtest" (DRT), "Deutsche Rechtschreibtest" (DERET), and "Hamburger Schreibprobe" (HSP). They are expensive to administer and cover word level and sentence level spelling errors where both words and sentences are manually tagged for predicted errors in predetermined words and texts that are either dictated to the child or elicited via pictures. Here, specific word material is requested to be written. By knowing the target word, the child is tested only on the forced vocabulary. Administration of these tests have been facilitated by providing online forms for tests (e.g. HSP-plus). "Gutschrift" by Löffler and Meyer-Schepers offers an online analysis tool based on a linguistic approach. "Lernserver" by Schönweiss at Universität Münster results in a diagnosis with personalised exercises. Additionally, an increasing number of schoolbook publishers are offering diagnosis online coupled with targeted learning material. A serious shortcoming with any of these types of tests, whether on paper or online, is the predetermined word and sentence material on which the child is tested. Manual tagging of spelling variations however is feasible here because of the known intended (target) words/text.

However, by far the most important limitation for predefined items is the limit on test-taking frequency. This problem may have been addressed in part by OLFA (Thomé and Thomé, 2010). While being somewhat text independent, manual annotation demands expertise by the teacher that makes its use somewhat difficult, not only for the teacher but particularly impossible for large-scale data processing.

The tool that is evaluated in this paper (building on previous work in this area (Berkling et al., 2011; Berkling and Lavalley, 2015) is able to automatically tag a number of detailed error categories and has been developed further to include more error categories and improve performance over the years since 2011 on the basis of spontaneously written text samples written by children from grades 2−8 (Lavalley et al., 2015). Detailed error categories refer to the underlying linguistic principles behind spelling error that is missing in a spell-checker and that will be explained in more detail in Section 5.