Writing Aid Dutch
Supporting Students' Writing Skills by Means of a String and Pattern Matching based Web Application

Margot D'Hertefelt, Lieve De Wachter and Serge Verlinde

Leuven Language Institute, KU Leuven, Dekenstraat 6, 3000 Leuven, Belgium

Keywords: Electronic Writing Aid, Higher Education Students, Writing Skills Development.

Abstract: Students at universities and colleges in Belgium and abroad often experience difficulties with writing (academic) texts in their native language (De Wachter and Heeren, 2011; Dugan and Polanski, 2006; Gray et al., 2005; Napolitano and Stent, 2009). This is reflected in many initiatives that are being developed specifically to support students' writing skills, among other the development of electronic writing assistance systems. Many of these systems are based on Natural Language Processing techniques, such as parsing. In this paper, we will argue that writing aids do not always have to make use of NLP techniques in order to analyze texts in a detailed and accurate way. We present an online writing aid, Writing Aid Dutch, which marks possible areas of concern in students' texts on three levels: (1) text structure and cohesion, (2) style and (3) spelling and provides users with individualized feedback. Writing Aid Dutch uses a lot of data and analyzes texts using complex queries and string matching techniques. Initial user experiences have been very positive so far. From February 2014 onwards, the effectiveness of the writing aid will be investigated in a one-group pre-post test design.

1 INTRODUCTION

Students at Flemish universities and colleges often have difficulties with writing, irrespective of the educational field they are in (Berckmoes and Rombouts, 2009; Berckmoes et al., 2010; Bonset, 2010; De Vries and Van der Westen, 2008; De Wachter and Heeren, 2011; Peters and Van Houtven, 2010). In 2011, a quantitative and qualitative needs analysis carried out among first year students of KU Leuven (Belgium) revealed that the most frequent writing problems of students are situated on the level of (1) text structure and cohesion, (2) style and, to a lesser extent, (3) spelling (De Wachter and Heeren, 2011). The results of this needs analysis are strikingly similar to those of previously conducted studies in Flanders as well as abroad.

The concern of students' poor writing skills is not confined to Belgium alone but is shared internationally and has already resulted in many initiatives offering writing support for students (Taylor and Paine, 1993; Gray et al., 2005; Dugan and Polanski, 2006; Graham and Perin, 2007). Among other things is the development of automatic and semi-automatic writing aids. Desktop applications such as SWAN (Scientific Writing AssistaNt, Kinnunen et al., 2012) or web applications such as the Language Tool Style and Grammar Checker (Naber, 2014) or Spell Check Plus (Nadashi and Sinclair, 2014) offer writing assistance to students who write at an L2 level or in their native language. These tools often use NLP techniques, such as a parser, to analyze the inserted texts in a detailed way.

Many of the writing assistance systems available today are able to provide students with useful and accurate feedback on different aspects of their text. However, despite the good intentions that they have, some of these writing assistance systems have some drawbacks as well. In the first place, the accuracy of the suggested feedback or corrections is not always satisfactory. Secondly, some of these writing aids, such as Scientific Writing AssistaNt, are rather time-consuming as students have to pass several 'stages' before receiving any feedback on their text. Moreover, SWAN provides the user with an overwhelming amount of information, which makes that he loses sight of the relevant feedback. This reduces the feeling of being responsible for your own writing product as well. Contrary to that, many

D'Hertefelt M., De Wachter L. and Verlinde S..
Writing Aid Dutch - Supporting Students' Writing Skills by Means of a String and Pattern Matching based Web Application.
DOI: 10.5220/0004948704860491
In *Proceedings of the 6th International Conference on Computer Supported Education* (CSEDU-2014), pages 486-491
ISBN: 978-989-758-020-8

web-based writing aids provide too limited feedback, which leaves the user frustrated and unsatisfied. Lastly, many writing aids concentrate too little on the writing process and do not encourage students' writing skills development, because they immediately suggest corrections (Napolitano and Stent, 2009).

In this paper, we present an online writing aid, the Writing Aid Dutch, a web application that responds to the strong need for effective writing support in Dutch. The writing aid analyzes texts, using string and pattern matching techniques to identify errors but also possible areas of concern in the submitted text. Based on the results of several needs analyses, the didactic purpose of the writing aid is to raise students' awareness on frequent writing problems that are situated on the level of (1) text structure and cohesion, (2) style and (3) spelling (Berckmoes et al., 2010; De Wachter and Heeren, 2011; Peters and Van Houtven, 2010). The writing aid does not correct and 'judge' students' writing mistakes, but marks them in the text and provides students with concise feedback, tips, examples and links to informative websites. Students can submit different genres of texts into the writing aid, such as a report, paper, essay, articles or master thesis.

In what follows, we will discuss the design and metrics of the writing aid after a short section on related work. We will then report some first user experiences and discuss future work, before we turn to our conclusions.

2 RELATED WORK

The development of Writing Aid Dutch fits in with an international trend of responding to students' writing problems with the development of electronic writing assistance systems. More specifically, it corresponds to the attention shift from product assessment to process-oriented support (Dale and Kilgarriff, 2011; Fontana et al, 2006; Gikandi et al., 2011). Writing assistance systems such as Amadeus (Fontana et al., 2006) or Helping Our Own (Dale and Kilgarriff, 2011) are specifically being developed to assist students throughout their writing process.

The underlying NLP techniques that these writing assistance systems use, however, differ from the data and string and pattern matching techniques that are implemented in Writing Aid Dutch. Apart from SOS-Frans ("SOS French") (Rymenams et al., 2012), a writing aid aimed at non-native speakers of French that has been developed at the same institute as Writing Aid Dutch, there is no knowledge of writing aids that do not make use of NLP techniques.

3 WRITING AID DUTCH

3.1 Interface

The interface of Writing Aid Dutch is simple and user-friendly: after students have copy-pasted or keyed in their text in the input field, they can click on three coloured buttons that each represent one of the three problem areas: (1) text structure and cohesion, (2) style and (3) spelling. These buttons are connected with arrows indicating the preferred order in which students should check the text. However, the student remains free to click on the button they prefer. As such, a learning path is suggested but students are free to determine their own pace in that they can choose which analyzed elements they want to look at first and when they want to take another step. The environment of Writing Aid Dutch is strongly user-controlled, seeing that our students are rather advanced learners and therefore do not need maximal guidance. Moreover, a system that is fully program-controlled would reduce the motivation of our students.

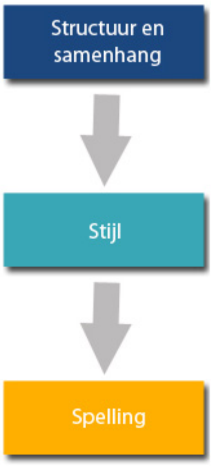
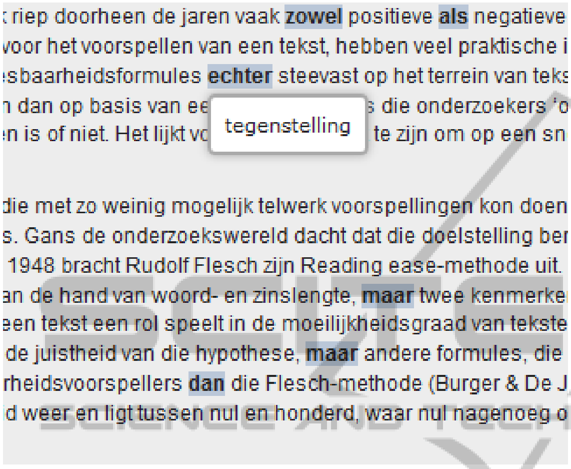


Figure 1: The three buttons 'Structure and cohesion', 'Style' and 'Spelling' on which students can click.

Considering that the writing tool is being developed for Dutch native speakers, feedback is in the form of general advice that is deliberately kept concise in order not to reduce students' motivation. For some of the text elements marked in the text, additional information is given in small pop-up screens that appear when the user scrolls over a highlighted text element, or in an extra field when students click on 'read more'. The illustration below gives a screenshot of text analysis and feedback for the use of structure words. When the user scrolls over a marked structure word, its meaning is provided in an extra pop-up field: in the illustration below, the meaning *tegenstelling* "contrast" is given for the structure word *echter* "however".



riep doorheen de jaren vaak zowel positieve als negatieve
voor het voorspellen van een tekst, hebben veel praktische i
sbaarheidsformules echter steevast op het terrein van teks
n dan op basis van ee | s die onderzoekers 'o
n is of niet. Het lijkt vo | te zijn om op een sn

die met zo weinig mogelijk telwerk voorspellingen kon doen
s. Gans de onderzoekswereld dacht dat die doelstelling ber
1948 bracht Rudolf Flesch zijn Reading ease-methode uit.
an de hand van woord- en zinslengte, maar twee kenmerke
een tekst een rol speelt in de moeilijkheidsgraad van tekste
de juistheid van die hypothese, maar andere formules, die
rheidsvoorspellers dan die Flesch-methode (Burger & De J
d weer en ligt tussen nul en honderd, waar nul nagenoeg o

Figure 2: Marking of structure words under 'Structure and cohesion'.

3.2 Metrics and Implementation

In each of the three levels, students can check specific textual elements or metrics that are related to it. In the following sections, the individual metrics of each level and the data involved will be described.

3.2.1 Level 1: Text Structure and Cohesion

In the level of text structure and cohesion the student can check (1) use of reference words, (2) use of structure words, (3) most frequent words of the text, (4) recurring sentence patterns, (5) sentence length and (6) paragraph length. More general statistics concerning text structure and cohesion, viz. the total number of words, sentences and paragraphs of the text are given as well. Lastly, the readability index (or complexity index) of the text is calculated.

Reference words and structure words are highlighted in the text by matching the text with lists of words. For the third metric, namely that of the most frequent content words of the text, the text is matched with a frequency list containing word forms of only content words. The word forms that are found in the text are lemmatized, and these lemmas are displayed to the student. As far as the next metric of recurring sentence patterns concerns, there is no specific measure. We have worked as follows: sentences that start with *de* "the", *het* "the", *een* "a", *die* "those", *dat* "that", *deze* "these", *dit* "this", *men* "one", *er* "there" point out to few variation in sentence construction. If more than two sentences in five start with these words, they are marked. This formula applies to other recurring words as well. For the last two metrics, sentence and paragraph length, a minimal and maximal boundary is set: sentences containing less than 8 words and more than 30 words are marked; the boundaries of the paragraphs are set at respectively 4 and 17 sentences per paragraph. For these two metrics, the average sentence and paragraph length is calculated and visualized through a small traffic sign, displaying a red ("too long/short sentences/paragraphs"), orange or ("possibly too long/short sentences/paragraphs) green ("sentence and paragraph length confirms to norm") light.

The readability index that is calculated is partly based on the Flesch-Douma formula, the readability formula based on Flesch (1948) but adapted to Dutch, which predicts a text's readability by taking into account word length, i.e. the number of syllables per word, and sentence length, i.e. the number of words per sentence. Despite a number of objections, such as the idea that long sentences are not always more complex than shorter ones (Jansen and Lentz, 2008), this formula has proven to be a reliable predictor of a text's readability and complexity. However, to make the formula even more accurate we have added word frequency, seeing that words that are highly frequent are more understandable than infrequent words. We use a frequency list consisting of word forms instead of lemmas.

3.2.2 Level 2: Style

The metrics distinguished in the second level are (1) use of passives, (3) use of nominalizations, (3) personal language use, (4) long-winded constructions, (5) informal and subjective words, (6) formal and archaic words, (7) vague words and (8) word combinations. For each of these metrics, Writing Aid Dutch checks whether the style of the inserted text is adapted to the required norm. Seeing that the students who use the writing aid come from different institutions (university or college) and, as a consequence, write in different text genres, the writing aid does not 'judge' the inserted text but provides the student with nuanced information about these different style requirements. Again, most of the metrics in this level are highlighted in the text by string and pattern matching.

3.2.3 Level 3: Spelling

The last level on which students can check their text is spelling, where typing mistakes and wrongly spelled words are marked by a spell-checker. The use of abbreviations is checked as well.

The implementation of the spell-checker has been (and still is) a labour-intensive work. The spell-checker is based on a word list containing over seven hundred thousand words forms that is still being completed. The database word list contains headwords supplied with linguistic information such as word class, article, plural form, past form, participle etc. In total, fifteen word classes are distinguished.

The spell-checker functions in various steps. The process starts by distinguishing every word separately, defining its boundaries by marking the spaces and punctuation marks and as such splitting up the sentence. After sentences are subdivided into separate words, occurrences of more or less fixed expressions are first of all being looked at. The database contains a list of these expressions, especially archaic phrases, which is matched with the text. A second step checks whether the remaining unrecognized and single words are in the word list. When this is not the case, the word will have to pass several conditions before it will be marked as wrong. In what follows, we will describe some of these conditions.

A first condition comprises combinations of numbers followed by a special character that are allowed in academic papers, for example “5°” or “10%”. A second condition refers to other symbols that may occur as well, such as Roman numbers like “I”, “IV” or “XI”. For the third and the fourth criterion, it is important to note that Dutch is a compound language in which words can very easily be composited. Compounds in Dutch are always written in one word or with a hyphen. The third selection criterion then concerns compound words that are written with a hyphen and consist of words that also exist on their own, for example a word such as *adjunct-directeur* “adjunct-director”. The fourth condition picks out compounds written without a hyphen. In this step, two functions are used to reduce the number of possibilities. A first one splits up a word, for example the word *strooizout* “road salt”, in the following manner:

```
s/trooizout
st/rooizout
str/ooizout
stro/oizout
stroo/izout
strooi/zout
```

The function stops when both queries give a valuable result, in this case *strooi* and *zout*. The minimal length for a word to be recognized is fixed at four characters, seeing that fewer characters resulted in too many false positives, i.e. words that do not exist but are nonetheless grammatically correct. A second function in this condition relates to the syntactic place that a particular word can have in a compound, namely in the beginning or at the end of the compound. This is statistically determined on the basis of the word list. For each syntactic option, frequency is calculated. For example, *achterover* “back” can never occur at the end of a compound but occurs, so far, a hundred and nine times in the beginning of a compound word, like in the verb *achteroverleunen* “to lean back”. In the fifth step of process, the spell-checker looks at a list containing named entities. When a word, then, still has not been found, the context is taken into account in order to check whether the word is part of a word group that has not been recognized as a fixed expression. Concretely, the context is limited to a span of four words left and right.

When a word still has not been recognized after these selection criteria, it will be marked red in the students’ text. However, a word can also be marked blue in the text. For these words, the spell-checker suggests an alternative form, based on the Levenshtein distance principle. This principle tries to alter one string into another string by making minimal changes, for example by changing or deleting one letter. The spell-checker is designed in a way that it is partly self-supportive. Unrecognized words automatically appear in a separate database, so that they, in the case of correct words, may be added later to the spell-check word list.

3.3 Comparison to Word Processing Software Such as Microsoft Word©

In Microsoft Word© grammar and spelling can be checked in a variety of languages, among which is also Dutch. A comparison between Microsoft Word© and Writing Aid Dutch seems therefore relevant. With regard to the computational implementation, language-specific information in Writing Aid Dutch cannot, unlike in Microsoft Word©, be considered as a rule set that is imported in the system. In the spell-checker of the writing aid, for example, many of the hard codes are only applicable to Dutch. An example is the following part of a code:

```
if(alleen_in_samenstelling($woord)
```

The part *alleen in samenstelling* “only in compound” relates to complex verbs in Dutch such as *tekeergaan* “to rant”. The part *tekeer* does not exist on its own but always occurs in combination with the verb *gaan* “to go”; as a consequence, *tekeer* will not be marked wrong because it is part of a complex verb. However, the codes that are used in Writing Aid Dutch to refer to its underlying databases can easily be adapted to other languages; only the databases itself will be different.

Because of the many complex and language-specific codes, the spell checker of Writing Aid Dutch is much more accurate and complete than the Dutch spell checker in Microsoft Word©. Checking grammar has never been a priority in the development of Writing Aid Dutch, seeing that its target audience are advanced native speakers of Dutch.

4 FUTURE WORK

4.1 Text Analysis on Content Level

At the moment, we are also experimenting with more content-oriented text analysis by categorizing certain words that appear in a student’s text into semantic fields. For this experiment we have used texts of KU Leuven students of Political Science, in which they had to compare two politicians. By identifying these words that express either similarity or difference in the text, the distribution of these two semantic categories is revealed, so that it can be investigated if they appear equally and at the right place in the text. Another experiment is the identification of academic words or more technical terminology in the text.

4.2 Effectiveness Analysis and Further User Study

From February 2014 onwards we will investigate the effectiveness of the writing aid in a quantitative and qualitative one-group design study. Despite the fact that such a design has minimal internal validity and no external validity (Sytsma, 2002), we have chosen this design because of time restrictions of the project. A within-subjects design does not require a placement test that cancels out possible differences in competencies between participants (de Smet et al., 2011). A total number of minimal 60 students of university as well as college institutions will be tested. On the one hand, effectiveness will be measured by rating texts written without and written with Writing Aid Dutch. On the other hand, students’ as well as teachers’ perception of the learning progress will be evaluated. The results of the effectiveness experiment will be available in June 2014.

A tool that is similar to Writing Aid Dutch, SOS-Frans, has been developed at the KU Leuven for French as a second and foreign language and turned out to be very effective, leading to fewer mistakes (Rymenams et al., 2012). Scientific Writing AssistaNt, reduced the lack of structure and semantic coherence in scientific papers (Kinnunen et al., 2012). Moreover, as teachers, we have already experienced noticeable progress in papers of students when they use Writing Aid Dutch. By analogy with similar writing aids and on the basis of our experiences, we hypothesize that the learning-process of students who use the Writing Aid Dutch will improve and that their writing products will be better.

As mentioned in Leakey (2011), the empirical data that result from quantitative research should ideally be completed with judgmental data. We have already gathered initial user experience by means of an online questionnaire filled in by 50 students. Next to students, 10 teachers of several faculties have reported their experiences in focus interviews. However, these data are not sufficient and we will carry out extra questionnaires and focus interviews with students and teachers as part of our effectiveness study.

5 CONCLUSIONS

In this paper, we have presented the Writing Aid Dutch. We have shown that the implementation of NLP techniques is not always a prerequisite for the development of appropriate computer-based support. Text analysis based on string and pattern matching techniques can be detailed, correct and fast. The writing aid (1) raises students’ awareness of frequent writing issues, (2) provides clear and individualized feedback, tips and examples, (3) focuses on the process, (4) has a simple and user-friendly design and (5) leads to less ‘shallow’ and repetitive correction work for lecturers. As a web application, the writing aid is a durable and partly self-supportive tool that can be adapted at any time.

REFERENCES

Berckmoes, D., Rombouts, H., 2009. Rapport verkennend onderzoek naar knelpunten taalvaardigheid in het hoger onderwijs. [Report preliminary investigation of higher education students' difficulties of literacy skills]. Antwerp: Linguapolis/University of Antwerp. Available at: <http://webh01.ua.ac.be/linguapolis/mom/Intern_rapport_verkennend_onderzoek_naar_knelpunten_taalvaardigheid_in_het_hoger_onderwijs-Monitoraat_op_maat.pdf> (Accessed 27 June 2012).

Berckmoes, D., Rombouts, H., Hertogs, K., 2010. Taalstimulering academisch Nederlands voor studenten aan de Universiteit Antwerpen. Monitoraat op maat. Rapport derde jaar, september 2008 – augustus 2009 (Language stimulation academic Dutch for students at the University of Antwerp. Report third year). Linguapolis/University of Antwerp.

Bonset, H., 2010. Nederlands in voortgezet en hoger onderwijs: hoe sluit dat aan? (Dutch in secondary and higher education: how does it match?). *Levende talen magazine*, 97(3), pp. 16-20.

Dale, R., Kilgarriff, A. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. *Proceedings of the 13th European Workshop on Natural Language Generation*, pp. 242–249.

De Smet, M. J. R., Broekkamp, H., Brand-Gruwel, S., Kirschner, P. A., 2011. Effects of electronic outlining on students' argumentative writing performance. *Journal of computer assisted learning*, 27, pp. 557-574.

De Vries, H., Van der Westen, W., 2008. Talige competentie in het hoger onderwijs (Linguistic competence in higher education). In: S. Vanhooren, A. Mottard, eds. 2008. *22ste conferentie Het Schoolvak Nederlands.* Gent: Academia Press. pp. 115-120.

De Wachter, L., Heeren, J., 2011. Taalvaardig aan de start. Een behoefteanalyse rond taalproblemen en remediëring van eerstejaarsstudenten aan de KULeuven (Entry-level academic language skills. A needs analysis of language problems and remedy of first year university students at the University of Leuven). Leuven Language Institute/University of Leuven. Retrieved from https://ilt.kuleuven.be/cursus/docs/Behoefteanalyse_TaalVaST.pdf.

Dugan, R. F. Jr., Polanski, V. G., 2006. Writing for computer science: a taxonomy of writing tasks and general advice. *Journal of Computing Sciences in Colleges*, 21(6), pp. 191-203.

Flesch. R., 1948. A new readability yardstick. *Journal of Applied Psychology*, 32, pp. 221-233.

Fontana, N. M., Caldeira, S. M. A., De Oliveira, L. C. F., Oliveira Jr., O.N. 2006. Computer assisted writing. Applications to English as a foreign language. *CALL*, 6(2), pp. 145-161.

Gikandi, J. W., Morrow, D., Davis, N. E. 2011. Online formative assessment in higher education: a review of the literature. *Computers and Education*, 57, pp. 2333-2351.

Graham, S., Perin, D., 2007. *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellent Education. Available at: http://www.all4ed.org/files/WritingNext.pdf (Accessed 12 September 2012).

Gray, E. F., Emerson, L., MacKay, B., 2005. Meeting the demands of the workplace: science students and written skills. *Journal of science education and technology*, 14(4), pp. 425-435.

Jansen, C., Lentz, L., 2008. Hoe begrijpelijk is mijn tekst? De opkomst, neergang en terugkeer van leesbaarheidsformules (How understandable is my text? The rise, downfall and comeback of readability formulas). Available at: http://www.kennislink.nl/publicaties/hoe-begrijpelijk-is-mijn-tekst (Accessed 1 January 2014).

Kinnunen, T., Leisma, H., Machunik, M., Kakkonen, T., Lebrun, J. L., 2012. SWAN – Scientific Writing AssistaNt. A tool for helping scholars to write reader-friendly manuscripts. *Proceedings of the 13th conference of the European chapter of the association for computational linguistics*, pp. 20-24.

Leakey, J., 2011. *Evaluating computer-assisted language learning. An integrated approach to effectiveness research in CALL.* Bern: Peter Lang.

Naber, D. 2014. Language Tool Style and Grammar Checker. Available at: www.languagetool.org (Accessed 1 January 2014).

Nadashi, T., Sinclair, S., 2001-2014. Spell Check Plus. Nadaclair Language Technologies. Available at: < http://spellcheckplus.com> (Accessed 1 January 2014).

Napolitano, D. M., Stent, A., 2009. TechWriter: an evolving system for writing assistance for advanced learners of English. *CALICO Journal,* 26(3), pp. 611-625.

Peters, E., Van Houtven, T., 2010. De weg naar materiaalontwikkeling is geplaveid met behoeftes (The way to material design is paved with needs). In: E. Peters, T. Van Houtven, eds. 2010. *Taalbeleid in het hoger onderwijs. De hype voorbij?*. Leuven: Acco. pp. 71-85.

Rymenams, S., Verlinde, S., Marx, S., Rosselle, M., Gerard, L., 2012. SOS-français: conception et évaluation d'un didacticiel d'aide à la rédaction interactif (SOS-French: development and evaluation of a didactic aid in an interactive environment). 3e Congrès Mondial de Linguistique Française. *SHS Web of Conferences*, 1, pp. 377-393.

Sytsma, S., 2002. The basics of experimental design. Available at: http://courses.washington.edu/bio480/Basics_of_Experimental_Design.pdf (Accesses 20 November 2012).

Taylor, H. G., Paine, K. M., 1993. An inter-disciplinary approach to the development of writing skills in computer science students. *Proceedings of the SIGCSE Technical Symposium on Computer Science Education*, pp. 274-278.