Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 397–404
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

An Exploratory Study into Automated Précis Grading

Orphée De Clercq[1] **and Senne Van Hoecke**[2]
[1]LT[3], Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Orphee.DeClercq@UGent.be
[2]TricS, Translation, Interpreting and Intercultural Studies, University of Antwerp
Grote Kauwenberg 18, 2000 Antwerp, Belgium
Senne.VanHoecke@uantwerpen.be

Abstract

Automated writing evaluation is a popular research field, but the main focus has been on evaluating argumentative essays. In this paper, we consider a different genre, namely précis texts. A précis is a written text that provides a coherent summary of main points of a spoken or written text. We present a corpus of English précis texts which all received a grade assigned by a highly-experienced English language teacher and were subsequently annotated following an exhaustive error typology. With this corpus we trained a machine learning model which relies on a number of linguistic, automatic summarization and AWE features. Our results reveal that this model is able to predict the grade of précis texts with only a moderate error margin.

Keywords: automated writing evaluation, natural language processing, learner corpus

1. Introduction

Research into technology to support writing has been an active research topic since the sixties (Page, 1966). Looking at present-day writing tools, a distinction can be made between automated essay scoring (AES) systems, which automatically assign a grade to a text; automated writing evaluation (AES) systems, which enable not only summative, but also formative feedback; and intelligent tutoring systems (ITS), which offer individualized instruction and feedback to students based on their needs (Allen et al., 2015).

Most of the currently existing systems, however, address English essay writing leading to a predominant focus on the argumentative text genre (Strobl et al., 2019), whereas in higher education academic writing skills are trained and tested by having students practicing a variety of text genres, culminating in a (bachelor's or master's) thesis.

This paper wishes to contribute research on a different text genre, summaries, and more specifically précis texts. A précis is a written text that provides a coherent summary of main points of a spoken or written text (Chan et al., 2015). It is usually formal, objective in tone and paraphrases the language of the original. In other words, précis are short formal texts in which content and paraphrasing play a significant role.

Though summarization is an effective strategy to promote and enhance learning and deep comprehension of texts (Graham and Herbert, 2010), it is seldom implemented by teachers in classrooms because the manual evaluation requires much time and effort (Crossley et al., 2019a). Nevertheless, a number of methods have been developed for automated summarization evaluation. Since the accurate capturing of the content is a crucial aspect (Li et al., 2018), most work in this respect has been done using Latent Semantic Analysis (Landauer et al., 1998) or by calculating n-gram overlap (Madnani et al., 2013) between summaries and source texts. On the other hand, linguistic features approximating vocabulary, syntax and cohesion using NLP processing on the source and summaries have also been used and have proven effective, both in combination with content n-gram overlap measures (Sladoljev-Agejev and Šnajder, 2017) and as stand-alone features (Crossley et al., 2019a).

The aim of this study is to present a corpus of English précis texts, written by first-year higher-education students whose mother tongue is Dutch. This writing genre is taught to students in their first year of higher education because it requires a combination of skills, i.e. listening, understanding, and writing, to be used at the same time, which presents a challenge for students and compels them to be clear beforehand on what they are expected to achieve. All texts received a grade assigned by a highly-experienced English language teacher and have been annotated following an exhaustive error typology.

We use this corpus to train a machine learning model which predicts a score between 0 and 20 (regression). As information sources we rely on a range of linguistic features and also include automatic summarization features and features derived from a popular AWE-tool. We perform two rounds of experiments: In the first round, all available features are used whereas in the second round only those features are incorporated which yield a significant correlation with the grades. In order to have an idea of the upper bound we each time distinguish between a system that only has access to automatically derived AWE-features and one that relies on the gold standard – annotated – error categories. Our results reveal that relying solely on the significant features already allows to construct a model that is able to predict the grade of précis texts with only a moderate error margin and that the difference between the fully automatic system and the upper bound is negligible.

The remainder of this paper is structured as follows: In the next section, we present a brief literature overview focusing on linguistic features which can be automatically derived from text. Section 3 elaborates on the corpus collection and

annotation with eleven distinct error categories. In Section 4 the method is explained in close detail and the results are presented in Section 5. We conclude this work and offer prospects for future research in Section 6.

2. Related work

This study attempts to map which linguistic features might be good grade predictors in the automatic grading of précis. While keeping in mind the typical writing style of this genre, certain linguistic features might be considered more important and might thus influence the grading process. As précis texts are generally condensed, yet logically ordered, versions of an original longer text, certain features dealing with lexicality, syntax and cohesion are expected to have a bigger influence on the comprehensibility and grading of the text.

It has been shown that vocabulary plays an important role in perceived text quality and comprehensibility (Pitler and Nenkova, 2008). Lexical features can be categorized and explained through the idea of lexical richness. Adhering to Read (2000), lexical richness consists of four closely related components, i.e. lexical density, lexical sophistication, lexical variation and number of errors in vocabulary use. All of these components focus on the use and presence of lexical items in a text. Kyle et al. (2018) studied the relation between lexical sophistication and perceived lexical proficiency in L1 and L2 in free writes, i.e. periods of constant writing disregarding idiomatic language, grammar or spelling often to overcome writer's block. They found positive correlations between indices concerning association strength, i.e. idiomatic word combinations, hypernymy, polysemy and Age of Exposure (AoE), indicating perceived lexical proficiency being related to the use of more specific, advanced and strongly related words. These findings are in line with another, preceding study by Kyle (Kyle, 2016) into the influence of lexical sophistication on scores on argumentative essays, both source-based and independent, written as a Test Of English as a Foreign Language (TOEFL).This study revealed a positive correlation between sophisticated vocabulary, e.g. less frequently used or more specific words or idiomatic word combinations, and scores of independent writing tasks.

In addition, the structure or syntax of a text is seen as an important contributor to its overall readability. Because longer sentences have proven to be more difficult to process than short ones (Graesser et al., 2004). This study adheres to Bulté and Housen (2012) and Kyle (2016) distinguishing two components of syntax: syntactic complexity and syntactic sophistication. Syntactic complexity entails the more countable aspects of syntax, e.g. the amount of subordination or average sentence length. Syntactic sophistication, or as Bulté and Housen (2012) refer to it: relative complexity, concerns indices related more to the reader's or writer's language proficiency. Syntactic sophistication is generally focused on Verb-Argument Constructions (VAC) and their relative corpus frequency, association strength, based on co-occurrences in corpora, and Age of Acquisition (AoA) scores (Kyle, 2016; Perfors et al., 2010). Most findings of L2 writing assessment research into syntactic influences (Danzak, 2011; Kyle, 2016) support the theories and results of readability research (Clercq and Hoste, 2016). Firstly, L2 writing assessment research has confirmed the importance of sentence length and, additionally, clause, phrase and T-Unit, i.e. a clause and its dependent clauses, length, even in grading (Danzak, 2011; Kyle, 2016). Secondly, similar to parse tree indices in readability, the number of dependents on a clause level, e.g. number of nominal subjects or number of direct objects per clause, have also been revealed to have a minor influence on grades of independent TOEFL essay writing tasks (Kyle, 2016).

Two major influencers of discourse that improve sentence relations and discourse structure comprehensibility are cohesion and coherence. Cohesion can be considered the glue that keeps a text together (Meyer, 2003). This 'glue' consists of linguistic features such as linking words, punctuation, pronouns, etc. (McNamara et al., 2015). It is a crucial element in a text as it aids the reader in following the author's train of thought and inter-textual links. Given the influence of cohesion and coherence on reading (Meyer, 2003), it would be expected that cohesion and coherence features have a significant influence on writing grades. However, research in the matter shows contradictory results. Although many agree that cohesion and coherence is key in qualitative writing, a number of studies (McNamara et al., 2010; Crossley and McNamara, 2011) reveal only minor to no correlations with grades in essay writing. McNamara et al. (2010) report a moderate, negative influence of using cohesive devices, connectives for example, in a text and a positive influence of coherence on judgments of writing quality. In a later study (Crossley et al., 2016b), the importance of overall text coherence in writing evaluation is confirmed, but also a positive influence of conjunctions, positive connectives, conjuncts and subordinators, contradicting the past findings. The authors explain this might be caused by a difference in the nature of the essays.

3. Corpus Collection and Annotation

3.1. Data

This study presents a corpus of 60 précis exam texts written in English by first-year undergraduate students from Applied Linguistics at Ghent University. All students are Belgian (Flemish), native speakers of Dutch and study English as a foreign language. The expected text length for the task was 190 words with an accepted deviation of 10% above or below; the students were allowed to use a monolingual dictionary of English; the full version of the text was delivered by an audio file that was played twice with a two-minute gap in between for the students to complete their notes; the students had one hour and fifteen minutes to write the précis after the audio was played twice. Please refer to the appendix for three example précis texts.

The original text discusses several theories concerning the origin of individualistic and collectivist civilisations, particularly a research conducted by the University of Virginia, which suggests that the agricultural difference of rice and wheat might be a possible cause for civilisations to tend towards collectivism or individualism respectively. It received a Flesch Reading Ease score of 44.8 and a Flesch-Kincaid Grade Level of 12.6. These scores can be interpreted as fairly difficult at the college level. The text it-

self was very accessible in terms of topic and language, except for a few expressions that might be difficult for first-year students to understand (i.e. 'analytic', 'falls at the first hurdle', 'influenza pandemics', 'mainstay', 'pooling of labour').
The exam took place at the end of a semester in which students had received their first higher education in English and had also received instructions on how to write précis texts in the framework of a general course on English structures. They had written and received individual feedback on one précis text before taking the exam.

3.2. Ratings

In an attempt to create an atmosphere of reasonable consensus on evaluating writing tasks and exams, the professors of the university's English department set out general band descriptors (viz. a rubric) for different text types. These descriptors list three grading categories (i.e. 9 or below, 10-13 and 14 and above), in which the characteristics are enumerated of a text belonging to that category. With the help of the agreed upon guidelines, different professors are more likely to assign the same approximate grade to a text before narrowing it down to its definitive mark, improving the odds of similar grading behaviour.
For the evaluation of précis texts, such band descriptors were also developed and used to grade all 60 précis texts by one expert evaluator. This was a native English professor who teaches writing classes and has more than twenty years of teaching experience. As illustrated in Figure 1, the broad majority of students (49 out of 60) has passed the exam. The minimum grade assigned was 6 (1 student) and the maximum 15 (2 students).

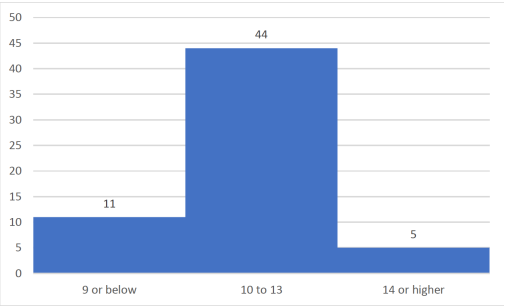

Figure 1: Distribution of précis grades according to the three grading categories

In a next phase, all texts were annotated based on an error typology which was directly derived from the previously mentioned rubric. This annotation work was carried out by a trained linguist as we did not get access to the original handwritten corrected copies. Eleven error categories were distinguished and most of them comprise various subcategories as described next:

- Grammar (Gram): This does not follow the grammatical rules of the English language (subtypes: verb form, subject-verb agreement, article-noun agreement, noun-adjective agreement, singular-plural, article, other).
- Lexicon (Lex): This does not follow the lexical rules of the English language (subtypes: wrong collocation, word non-existent, wrong preposition, other).
- Spelling, typos and capitalisation (Spel_Typ): This either does not follow the rules of spelling and punctuation of the English language or is a typo (subtypes: spelling, punctuation, capitalisation, compound, hyphen, other).
- Style and Register (Style_reg): The choice of words and phrasing does not meet the expectations of the present text type (subtypes: long sentence, short sentence, inappropriate words or phrases, register, extensive use of passive voice, repetition of words, other).
- Coherence/Cohesion (Coher): There is something wrong with the coherence of the text (subtypes: strange or missing conjunction, confusing sentence, inconsistent use of terms, other).
- Interference (Inter): The use of this word or structure was influenced by the writer's native language and is considered incorrect when used in English (subtypes: Dutch phrasal verb, word translated from Dutch, false friend).
- Paragraphing (Parg): Either paragraphing is clear and effective throughout the text or no attempt to paragraph was made or paragraphs bear little relation to the units of content.
- Paraphrasing (Para): The writer has not used the same wording as the original text (subtypes: minimal, attempt made or consistent).
- Content (Cont): A piece of information is either missing or interpreted in the wrong way (minimum one and maximum five key content points).
- Length (Len): The text either exceeds or falls short of the expected word count for the writing task.
- Other: This error cannot be placed into one of the other major categories

In order to validate whether these error categories are good indicators of the assigned grades, correlations were calculated with the occurrence of errors as presented in Table 1. These numbers reveal that eight out of the eleven error types correlate with the grades. The error type with the strongest positive correlation is the level of paraphrasing (r: 0.499) and the number of spelling, typo and capitalization errors reveal the strongest negative correlation (r: -0.485).

4. Method

Following the work of Crossley et al. (2019a), our primary objective is to automatically assign a grade to a précis text. Whilst in their work they only rely on linguistic features derived from various NLP tools, we also have a look at other features based on research in the field of automatic summarization and error indications by an automated writing evaluation tool. Correlations between the grades and the derived features are first calculated after which machine learning experiments are carried out.

Error type	Correlation	Sig.
Gram	-0.408772300	**
Lex	-0.209359806	*
Spel_Typ	**-0.485355734**	***
Style_reg	-0.249561342	
Coher	-0.329043833	*
Inter	-0.276819640	*
Parg	0.061274755	
Para	**0.499260595**	***
Content	0.266497823	*
Len	0.073154227	
Other	-0.369522104	**

Table 1: Correlations between the error categories and the assigned grades

4.1. Information sources

All 60 précis texts were processed in order to derive a number of features.

4.1.1. Linguistic features

In order to extract linguistic features a number of publicly available feature extraction tools were selected because of their availability, simplicity and user-friendliness. These tools allow to derive information about the lexical sophistication, lexical diversity, syntactic complexity and cohesion of the précis texts amounting to 570 features in total (Table 2).

Linguistic features	# feats
Traditional	4
Lexical	157
Syntax	253
Coherence & Cohesion	156
TOTAL	570

Table 2: Number of linguistic features extracted per précis

SiNLP[1] (Crossley et al., 2014) was used to extract four traditional readability features, namely mean word length, mean sentence length, number of paragraphs and number of sentences.

Next, TAALES[2] (Kyle and Crossley, 2015) and TAALED[3] were employed to extract lexical features. In this respect two major categories can be distinguished: lexical sophistication and lexical diversity and density features.

The lexical sophistication features are based on word frequencies, word ranges, psycholinguistic word information, polysemy and hypernymy and n-grams. Both word frequencies and word ranges are calculated based on various background corpora, namely the Brown Corpus, the Thorndike-Lorge Corpus, the Academic Corpus (Academic Word List, AWL) the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). As the dataset consists of student-written examination texts and the focus of the examination itself lies on academic language competences, the word frequency indices that are based on COCA have been limited to only academic word frequencies, i.e. COCA Academic sub-corpus. The n-gram features measure the number of multi-word combinations compared to their presence in the included corpora.

Regarding the lexical density features this study includes three features for the simple and three for the root TTR (type token ratio) with the TTRs being separated on content or lexical words, function words and all words.

In a third step, TAASSC[4] (Kyle, 2016) was used to extract syntax features. These consist of features concerning relative complexity, also called cognitive complexity or syntactic sophistication, and those concerning absolute complexity, i.e. syntactic complexity. Specific for the tool we used here, is that it also incorporates the Second Language Syntactic Complexity Analyzer (L2SCA), a system specifically trained on college-level L2 writing and developed for the analysis of syntactic complexity L2 writing (Lu, 2010). It includes fourteen features based on syntactic indices used in L2 research.

Finally, the tool TAACO[5] (Crossley et al., 2019b) was used for coherence and cohesion features. A division can be made between local cohesion, i.e. cohesion across sentences, global cohesion, i.e. cohesion across paragraphs, and overall cohesion, i.e. cohesion in the entire piece. Research has shown that the use of cohesive devices correlates with perceived writing quality in L2 writing. However, there is some discussion as to how cohesion influences ratings of writing. Global cohesion indices appear to be predictive of essay quality more frequently and consequently than local indices. At least two studies (Crossley et al., 2016b; Crossley and McNamara, 2010) report positive correlations between global cohesive indices, e.g. similarity and overlap between paragraphs, and overall writing quality and organization judgments. With regard to text cohesion, previous research has yielded some diverse results. Some studies (Crossley and McNamara, 2010; Crossley and McNamara, 2011; McNamara et al., 2010) report a negative to a non-existing relation between essay quality and text cohesive indices. Crossley et al. (Crossley et al., 2016b) on the other hand show that text cohesive indices are predictive of writing score and organization judgment.

4.1.2. Summarization features

We calculate two automatic measures that have been used in the past to compare human with computer summaries. Both metrics calculate some sort of overlap between a gold and hypothesis version of the same text. We had access to the transcript of the original audio file that was played two times during the examination and a model précis that was composed by the expert evaluator (see Figure 3 in the Appendix).

The BiLingual Evaluation Understudy or BLEU metric (Papineni et al., 2002) is a metric that has extensively been used to score machine translation output. This precision-based metric computes n-gram overlap between the précis text written by the students and the original text. This metric will thus indicate how many of the words and phrases are directly coming from the original text.

[1] https://www.linguisticanalysistools.org/sinlp.html
[2] https://www.linguisticanalysistools.org/taales.html
[3] https://www.linguisticanalysistools.org/taaled.html
[4] https://www.linguisticanalysistools.org/taassc.html
[5] https://www.linguisticanalysistools.org/taaco.html

Next, the Recall-Oriented Understudy for Gisting Evaluation or ROUGE (Lin, 2004) was also calculated between the student summaries and the model précis. This metric is recall-based and measures the lexical and phrasal overlap between the student précis and the model précis. It is calculated in various flavours: ROUGE-L looks at the longest common subsequence, ROUGE-1 refers to the overlap of unigrams and ROUGE-2 the overlap of bigrams.
Note that an essential skill for writing good précis texts is paraphrasing and that the degree of paraphrasing yielded the highest positive correlation with the assigned grades (r: 0.499).

4.1.3. Writing evaluation features

All student précis texts were processed with the freely available part of Grammarly[6]. We deliberately chose this version as it is open to everyone which means that students could also have easy access to this. According to Grammarly, the online grammar checker scans your text for different mistakes, ranging from typos to sentence structure problems. It is based on a sophisticated AI that does not only rely on rules but also takes into account the context when making corrections or suggestions (Grammarly Inc).

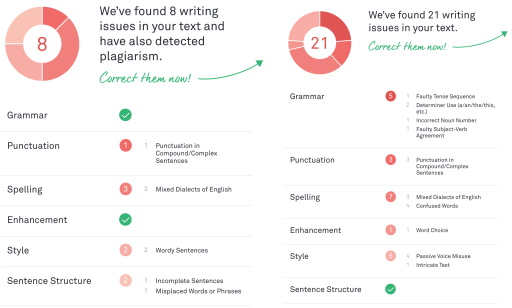

Figure 2: Grammarly error feedback on two student précis texts: highest grade (left) versus lowest grade (right).

Figure 2 depicts two output screenshots we received for two précis texts written by students. The first one is a text that received the highest grade (15) and the second one a text which received the lowest grade (6). As can be derived, the Grammarly software distinguishes between six error types: grammar, punctuation, spelling, enhancement, style and sentence structure. The same error types will be used as features.

4.2. Experimental setup

Ten-fold cross-validation machine learning experiments were set up in order to test whether it is actually possible to train a model that can learn to predict the précis grades using the information sources presented in the previous sections. To this purpose the dataset of 60 précis texts were split into ten folds each containing 54 texts for training and six texts for testing. Two different flavours were tested: one where all features were combined and one where only the significant features were retained. We also had two different setups regarding the AWE information. In a first setup only the automatically derived features were used (i.e. the error types coming from Grammarly) and in the second setup the gold standard error categories were added to the features. This latter setup should allow us to find the upper bound.
All experiments were conducted using support vector machines (SVMs), and more specifically the LibSVM[7] implementation which supports support vector regression. The output was evaluated by each time calculating the Root-Mean-Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - x_i)^2}$$

in which X_i is the prediction, x_i the response value, that is the correct value, for the regression task at hand, and m is the number of texts for which a prediction is made. RMSE computes the differences between the predicted and observed values, meaning it measures the overall accuracy of the model, the lower the RMSE, the better.

5. Results

Of the 570 assessed linguistic features, only 33 features revealed significant correlations (see Table 3). Regarding the traditional features none correlated. In line with our expectations, a moderate number of lexical features showed significant correlations with the assigned grades, namely 17 out of the 157 lexical indices. The syntactic feature category yielded 13 significant features out of 253 syntactic features. Of all 156 coherence and cohesion features, only three showed correlation with the assigned grades. This study expected cohesion and coherence features to show a number of significant correlations. Previous grading research (McNamara et al., 2010; Crossley and McNamara, 2011), however, has risen doubts about the importance of cohesion measures in automatic grading, or possibly the difficulty of computing a text' cohesion. This study supports these earlier findings.
Contrary to our expectations, both automatic summarization metrics did not yield a significant correlation with the assigned grades as illustrated in Table 4. For BLEU the précis texts were compared to the original text to check for literal repetitions of words or phrases, but apparently n-gram overlap is not a good indicator for the assigned grade. The same goes for the ROUGE metrics, where the précis texts were compared to the model précis written by the teacher, it seems that more or less strict adherence to this model précis does not influence the actual grading.
Table 5 presents which Grammarly features correlate with the assigned grades. The three significant error types were to be expected as the strongest correlations between the grades and the annotated error categories were also Spel_Typ and Grammar (see Table 1).
Finally, table 6 presents the results of the machine learning experiments. These reveal that relying only on the significant features allows for a model that is able to predict the correct grade with an error margin of around 1.6. The difference between the setup with automatically-derived features and gold standard error categories is marginal. What

[6]https://www.grammarly.com/grammar-check

[7]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Linguistic features	Correlation	Sig.
TL_freq_CW	-0.262253521	*
Brown_freq_CW	-0.257961755	*
Brown_freq_CW_Log	-0.259404762	*
All_AWL_normed	0.307902066	*
MRC_Meaningfulness_AW	-0.293563784	*
MRC_Concreteness_FW	-0.263572020	*
MRC_Imageability_FW	-0.295273905	*
MRC_Meaningfulness_FW	-0.286744357	*
Kuperman_AoA_AW	0.284283047	*
Kuperman_AoA_CW	0.366821563	**
aoe_inverse_linear_regr_slope	0.256143648	*
aoe_inflection_point_polyn	0.269953100	*
poly_verb	-0.274251029	*
hyper_verb_noun_Sav_P1	0.254891017	*
hyper_verb_noun_Sav_Pav	0.264462225	*
COCA_academic_bi_MI2	0.294925243	*
COCA_academic_bi_DP	0.275411117	*
av_nsubj_deps	0.256142414	*
av_nominal_deps_NN	0.308795101	*
av_nsubj_deps_NN	0.363631367	**
prep_all_nom_deps_struct	0.261772122	*
rcmod_all_nom_deps_NN_struct	-0.268265654	*
advmod_all_nom_deps_NN_struct	0.264554662	*
det_nsubj_deps_NN_struct	0.276806550	*
rcmod_pobj_deps_struct	-0.294770840	*
rcmod_pobj_deps_NN_struct	-0.288811859	*
ccomp_per_cl	-0.296615064	*
tmod_per_cl	0.270964840	*
MLT	0.267830037	*
MLC	0.353782090	**
all_logical	0.268877691	*
positive_logical	0.274558684	*
positive_intentional	-0.427045707	***

Table 3: Correlations between the linguistic features and the assigned grades

Metric	Correlation	Sig.
BLEU	**0.210911316**	
ROUGE-L (prec.)	0.054125372	
ROUGE-L (rec.)	0.117912667	
ROUGE-L (F)	0.091735391	
ROUGE-1 (prec.)	0.138674577	
ROUGE-1 (rec.)	0.186920118	
ROUGE-1 (F)	**0.163359814**	
ROUGE-2 (prec.)	0.096322671	
ROUGE-2 (rec.)	0.174263131	
ROUGE-2 (F)	0.139356820	

Table 4: Correlations between the summarization features and the assigned grades

draws the attention is that when including all features, the gold error categories seem to decrease the performance of the system, whereas the opposite is true when only the significant features are included. This latter approach achieves the best overall, an RMSE of 1.562.

Grammarly	Correlation	Sig.
Grammar	-0.364881214	**
Punct	-0.262258198	*
Spelling	-0.416647974	***

Table 5: Correlations between the Grammarly features and the assigned grades

	All features	Significant features
Automatic	**1.83319**	1.62804
Gold	1.90059	**1.56156**

Table 6: Results, expressed in RMSE, of the 10-fold cv experiments

6. Conclusion

In this paper, we presented a corpus of English précis texts, written by first-year higher-education students whose mother tongue is Dutch. All texts received a grade assigned by a highly-experienced English language teacher and were annotated following an exhaustive error typology. The corpus is available for research purposes[8].

We used this corpus to train a machine learning model to perform a regression task. As information sources we relied on a range of linguistic features and also included automatic summarization features and features derived from the popular AWE-tool Grammarly. By calculating the correlations between all these different features and the assigned grades, we distinguished between two setups and found that relying solely on the significant features already allows to construct a model that is able to predict the grade of précis text with only a moderate error margin. In a final step, a version of the system which relied on gold-standard writing errors was compared to one including the automatically derived error types from Grammarly. The differences in performance between both setups are negligible.

The presented machine learning experiments are limited as only a corpus of 60 précis texts was available for training. However, these first results are promising and, in future work, we would like to further corroborate these findings on larger datasets and also explore whether and how we can incorporate and leverage existing data. We also envisage to collect more text material, encompassing various writing tasks in different languages.

Acknowledgements

We want to thank the first-years students and David Chan at the Ghent University Department of Translation, Interpreting and Communication for providing us with the graded corpus of précis texts. We also wish to express our gratitude to Guillaume Goemanne for annotating the corpus with the eleven error categories and to the reviewers for their valuable suggestions.

Appendix: précis examples

Figures 3 and 4 present three different précis texts. The model précis that was composed by the English professor and two précis texts written by students.

[8] https://www.lt3.ugent.be/resources/

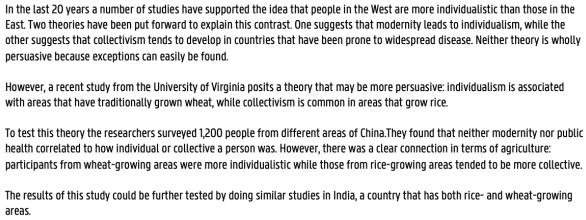

Figure 3: Model précis text

Psychological research over the last two decades has confirmed that there is a clear difference in the way Orientals and Occidentals think. Westerners tend to have a more individual, analytic and abstract mental life, whereas Asians live in collective existence. The research includes three hypotheses.

The first hypothesis presumes that modernisation promotes individualism. However, Japan, a western-like industrialised country, has a rather collective outlook. Second, in regions with a high grade of infectious diseases, people are cautious when meeting strangers. As a result, groups tend to turn inwards. Thirdly, Thomas Telhalm assumes that the crucial difference between West and East is agricultural. Whereas the western stable crop is wheat, Asians like rice.

The collective outlook would dominate society's cultural behaviour, which is testable by comparing agriculture in one country. In China, for example, the contradiction between collectivism and individualism nor corresponds to the place of origin, nor to the public health level. Researchers concluded that there tends to be more individualism in wheat-growing areas.

In conclusion, it is clear that different psychologies between East and West as a consequence of agriculture need further exploration.

As it has already been known for centauries, Westerns and Easterns don't think alike. It is believed that Westerns behave more individually, whereas Easterns are more collective thinkers. There are three hypothesis to explain these differences.

Initially, modernisation was regarded as a reason for individualism. Currently, Japan has asserted the opposite. Modernisation and collective thinking go perfectly together.

An other explanation for individualism is public health. Infectious diseases make people more cautious about making contact with others, which leads to antisocial behaviour. This theory does not give a full explanation. Even thought Europe has had a plague in the past, Asia too has dealt with infectious diseases.

A researcher in Virginia has come with an other hypothesis. His study shows that agricultural tradition influences the way people think and not their origin. In the West, people easily grow wheat, while in the East farmers work twice as hard on their rice fields. That is why they learn to work in teams. This does not mean everyone in the East thinks collectively. In some regions in the East where wheat is cultivated, the inhabitants tend to be more individual.

Figure 4: The précis texts which received the highest (left) and lowest (right) grade.

7. Bibliographical References

Allen, L. K., Jacovina, M. E., and McNamara, D. S. (2015). Computer-Based Writing Instruction. In Charles A. MacArthur, et al., editors, *Handbook for Writing Research, Second Edition*, pages 316–329. Guilford Publications.

Bulté, B. and Housen, A. (2012). Defining and operationalising L2 complexity. *Language Learning Language Teaching*, 32:21–46.

Chan, D., Jooken, L., and Robberecht, P. (2015). *Writing in English: advanced English writing skills for Dutch speakers*. Academia Press, Ghent.

Clercq, O. D. and Hoste, V. (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*, 42(3):457–490.

Crossley, S. A. and McNamara, D. S. (2010). Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, page 32(32).

Crossley, S. A. and McNamara, D. S. (2011). Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1231–1236.

Crossley, S. A., Allen, L. K., Kyle, K., and McNamara, D. (2014). Analyzing discourse processing using a simple natural language processing tool (SiNLP). *Discourse Processes*, 51(5-6):511–534.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016b). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Crossley, S. A., Kim, M., Allen, L., and McNamara, D. (2019a). Automated summarization evaluation (ase) using natural language processing tools. In Seiji Isotani, et al., editors, *Artificial Intelligence in Education*, pages 84–95, Cham. Springer International Publishing.

Crossley, S. A., Kyle, K., and Dascalu, M. (2019b). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27, Feb.

Danzak, R. L. (2011). The integration of lexical, syntactic, and discourse features in bilingual adolescents' writing: an exploratory approach. *Language, Speech, and Hearing Services in Schools*, 42(4):491–505.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

Graham, S. and Herbert, M., (2010). *Writing to Read: Evidence for How Writing Can Improve Reading: A Carnegie Corporation Time to Act Report*. Alliance for Excellent Education, Washington.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Kyle, K., Crossley, S. A., and Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, 50(3):1030–1046.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Doctoral dissertation, Georgia State University.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Li, H., Cai, Z., and Graesser, A. C. (2018). Computerized summary scoring: crowdsourcing-based latent semantic analysis. *Behavior Research Methods*, 50(5):2144–2161, Oct.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lu, X. (2010). Automatic analysis of syntactic complex-

ity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Madnani, N., Burstein, J., Sabatini, J., and O'Reilly, T. (2013). Automated scoring of a summary-writing task designed to measure reading comprehension. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168. Association for Computational Linguistics.

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic Features of Writing Quality. *Written Communication*, 27(1):57–86.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Meyer, B. J. F. (2003). Text Coherence and Readability. *Topics in Language Disorders*, 23(3):204–224.

Page, E. B. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, 47(5):238–243.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Perfors, A., Tenenbaum, J. B., and Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37(3):607–642.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 186–195, Honolulu, Hawaii.

Read, J. (2000). *Assessing Vocabulary*. Cambridge Language Assessment. Cambridge University Press.

Sladoljev-Agejev, T. and Šnajder, J. (2017). Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in L2. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 181–186. Asian Federation of Natural Language Processing.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., and Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers Education*, 131:33 – 48.