Investigating Productive and Receptive Knowledge: A Profile for Second Language Learning

Leonardo Zilio **Rodrigo Wilkens** **Cédrick Fairon**
Centre de Traitement Automatique du Langage (Cental)
Université Catholique de Louvain (UCL), Belgium
{leonardo.zilio,rodrigo.wilkens,cedrick.fairon}@uclouvain.be

Abstract

The literature frequently addresses the differences in receptive and productive vocabulary, but grammar is often left unacknowledged in second language acquisition studies. In this paper, we used two corpora to investigate the divergences in the behavior of pedagogically relevant grammatical structures in reception and production texts. We further improved the divergence scores observed in this investigation by setting a polarity to them that indicates whether there is overuse or underuse of a grammatical structure by language learners. This led to the compilation of a language profile that was later combined with vocabulary and readability features for classifying reception and production texts in three classes: beginner, intermediate, and advanced. The results of the automatic classification task in both production (0.872 of F-measure) and reception (0.942 of F-measure) were comparable to the current state of the art. We also attempted to automatically attribute a score to texts produced by learners, and the correlation results were encouraging, but there is still a good amount of room for improvement in this task. The developed language profile will serve as input for a system that helps language learners to activate more of their passive knowledge in writing texts.

Title and Abstract in Portuguese

Investigação de Conhecimento Produtivo e Receptivo: Um perfil de aprendizado de segunda língua

A literatura de aquisição de segunda língua frequentemente aborda diferenças no vocabulário produtivo (que um aprendiz consegue utilizar em textos) e receptivo (que um aprendiz consegue entender em textos) de um aprendiz de língua, mas as estruturas gramaticais normalmente não são investigadas nesse nível. Neste trabalho, usamos dois corpora para investigar as divergências nas ocorrências de estruturas gramaticais pedagogicamente relevantes em textos de produção e recepção. Os valores de divergência observados foram incrementados com a atribuição de uma polaridade que indica um sobreuso ou subuso de uma estrutura gramatical por aprendizes de língua. Essa investigação levou à compilação de um perfil de linguagem voltado ao aprendizado de segunda língua que posteriormente foi combinado com informações lexicais e de legibilidade para uma classificação de textos de produção e recepção em três categorias: iniciante, intermediário e avançado. Os resultados da classificação automática tanto de textos de produção (medida F de 0,872) quanto de recepção (medida F de 0,942) foram comparáveis ao atual estado da arte. Também realizamos um experimento para atribuir automaticamente uma nota aos textos produzidos por aprendizes, e os resultados da correlação foram encorajadores, mas mostram que ainda há muitas lacunas a serem supridas e questões em aberto para a realização da tarefa, especialmente no que diz respeito à subjetividade envolvida na atribuição de notas. O perfil de linguagem apresentado servirá como base para um sistema de auxílio à ativação de conhecimento receptivo na escritura de textos.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3467–3478
Santa Fe, New Mexico, USA, August 20-26, 2018.

1 Introduction

Second Language Acquisition (SLA) involves a series of different linguistic knowledge that a second language learner has to deal with, such as, but obviously not limited to, vocabulary and syntax of the second language. Besides the linguistic levels, another important topic for SLA is the knowledge activation capacity, i.e., how much from the learner's knowledge can be activated during a productive task. These are all aspects that help conform the SLA process, and are factors that get included in the evaluation of second language learners in terms of their capabilities of performance in a second language environment.

In the literature, it is possible to find many studies discussing how the passive/receptive vocabulary knowledge relates to the active/productive vocabulary knowledge, be it in a native or second language setting (e.g. Morgan and Oberdeck (1930), Meara (1990), Laufer (1998), Fan (2000), Schmitt (2008), Pignot-Shahov (2012)). In those studies, it is agreed that the passive vocabulary knowledge is larger than the active one. However, none of those studies devote some attention to grammar, so that it is not debated how grammar works in terms of productive and receptive knowledge.

In the case of second language learning, there is also the adherence to a commonly used framework, namely the Common European Framework of Reference for Languages (CEFR) (Verhelst et al., 2009), which divides and describes the communicative goals that should be achieved by an idealized second language learner, no matter in which language or from which native language, in each of the levels. The levels scale from A to C, and there are two sublevels each: A1, A2, B1, B2, C1, and C2. The CEFR also conceives the existence of a second subdivision, such as A1-, A1, and A1+, so that, if needed, the sublevels can be further specified. As such, if one intends to deal with SLA, there is a lot of ground to cover in terms of knowledge that is spread along all the different levels and subclassifications.

One of the ways of dealing with this spread knowledge is by organizing language profiles. The methods for characterizing a language profile use as basis the observation of different groups of people that use language in distinct ways, but that have some fundamental common traces when compared to other groups of speakers of the same language. The whole of these common traces can be called a language profile (Argamon et al., 2009). Two examples of profiles that deal with a specific linguistic level for language learning are the English Vocabulary Profile and the English Grammar Profile[1].

In this study, we are interested in the observation of receptive and productive knowledge present in second-language-related corpora to draw information about the differences between expected receptive and actual productive knowledge. We compare texts designed as input for SLA and the actual output of learners of English to see where the divergence between both of these types of tasks lies. Since vocabulary has already been studied in various works, here we will be using vocabulary as well, but we will deal especially with grammar, and we aim at automatically generating a language profile that can model the relationship between production and reception in SLA. We further hypothesize that, by adding grammatical information, we can improve the modeling of SLA in terms of productive and receptive knowledge.

The language profile developed in this study will also be used to help learners in the task of producing written texts that use the most from the learner's passive knowledge. For achieving this goal, we describe the relation between reception texts and production texts in terms of their divergence, and later we extrinsically evaluate our model, by applying the information drawn for our language profile to a text classification task. As an associated task, we also delve into the automatic evaluation of texts produced by language learners, by correlating scores generated by a classification algorithm to the actual scores given by professional evaluators.

This paper is divided as follows: in Section 2, we briefly contextualize the task addressed in this study and how we are going to deal with it; in Section 3, we describe the corpora that we used as representation of reception and production texts and their automatic annotation with pedagogically relevant grammatical information; we then explain the grammatical profiling that was carried out in Section 4; Sections 5 and 6 are dedicated to the experiments that we conducted using our language profile, these two sections present

[1] Both the EVP and the EGP are available at http://www.englishprofile.org/.

the methodology used for each experiment and also discuss their results; finally, we sum up our findings and report some future work in Section 7.

2 The Task at Hand

In this paper, we are dealing with the contrast between texts produced by learners and texts that are used as input for learners. Our main goal is to draw from this type of data a language profile that can be used to help language learners to produce texts that use the most of their passive knowledge. In this paper, we deal with texts written in English, but the methodology could be applied to other languages, provided there are similar resources.

The contrastive study of distributions was successfully used to identify characteristics that are relevant to text profiling (Biber, 1991). In a complementary manner, entropy-based approaches have been employed to identify similarities (and dissimilarities) among text sets (Dagan et al., 1997; Oakes, 2008) and to study vocabulary variation (Oakes and Farrow, 2007). In this respect, there are different divergence equations that can be employed for evaluating entropy, and one that stands out is the Kullback–Leibler divergence (Kullback and Leibler, 1951; Kullback, 1997), which is also called relative entropy. This divergence is defined in Equation 1, where P and Q are the probability distribution of two text sets. Despite its usefulness, it has a known problem: a lack of symmetry, which makes it hard to use in terms of a distance measure. In that sense, the Jensen-Shannon divergence (Endres and Schindelin, 2003; Österreicher and Vajda, 2003; Fuglede and Topsoe, 2004), presented in Equation 2, averages the divergence of the interchange of the distributions taking into account the average of the distribution P and Q, as shown in Equation 3. The Jensen-Shannon divergence solves the lack of symmetry, so that the result can be understood as a measure of distance, which allowed us to use it to observe the difference between reception and production.

$$D_{KL}\left(P||Q\right) = -\sum_{i} P(i) log\left(\frac{Q(i)}{P(i)}\right) \tag{1}$$

$$D_{JS}\left(P||Q\right) = \frac{1}{2} D_{KL}\left(P||M\right) + \frac{1}{2} D_{KL}\left(Q||M\right) \tag{2}$$

$$M = \frac{1}{2}\left(P + M\right) \tag{3}$$

For achieving the objectives described in this paper, we used two corpora: one made up of texts that are used as input in language learning tasks, and the other one composed of texts that were produced by language learners. Different parts of these corpora were used for deriving our language profile, which was based on a Jensen-Shannon Divergence model, and for extrinsically testing our profile in different tasks related to the evaluation of texts used in language learning.

By observing the divergence between production and reception and generating a language profile, we can also create a tool that guides language learners to potentially use more of their passive knowledge in written production tasks. The idea here is to match a learner's written text to the patterns found in reception texts, so that we can provide cues for the learner on how the text could be improved.

3 Corpora

In this study, we conducted several separate experiments that were related to the goal of developing a language profile that takes into account both sides of written language learning and applying this profile to classify and evaluate written texts in accordance to a given language level. The experiments had as basis two corpora: one corpus containing texts used as input for language learning, and another one containing texts that were produced by language learners. In the next two subsections, we describe the sources we used for both corpora. Since we used them differently in each of the experiments[2], here we provide only a general description, leaving the more fine details to be described together with the

[2]It is important to make it clear that we did not mix the texts used for creating the profile with those that were posteriorly used for extrinsically testing it.

experiments. At the end, we describe the automatic annotation that was applied to both corpora for the experiments.

3.1 Reception Corpus

For observing the type of linguistic information, in terms of vocabulary and grammar, that is present in the input for a language learning task, we combined texts from two different sources: the Breaking News English Lessons (BNE) (Banville, 2009), and Altissia's animation database for the English course[3].

The BNE corpus consists of texts written for learners of English that are distributed along the Common European Framework of Reference for Languages (CEFR) levels, albeit using a level classification from 0 to 7 (that covers the levels A2 to C2). It contains documents that range from 120 words on the easiest level to 250 words on the higher levels. Since the CEFR levels are mixed in some of the texts, what we did was to consider a division of the documents into a beginner class (for those addressing the A2 level), an intermediate class (for those addressing B1 and B2 levels), and an advanced class (for those addressing C1 and C2 levels). As such, levels 1 and 2 of the BNE were allocated to the beginner class; levels 4 and 5 composed the intermediate class; and levels 6 and 7 formed the advanced class. Level 3 was discarded for being a hybrid between the levels A2 and B1, and level 0 was discarded for balancing reasons. The final BNE corpus that we used contains 930 documents and 194,207 tokens.

Altissia's animation database is composed of transcribed texts from short animation sequences, comprising dialogs and narratives that are used as input for learners on an online language learning platform. The transcriptions are all classified according to the CEFR, so we put together texts from levels A1 and A2 (for the beginner class), B1 and B2 (for the intermediate class), and C1 and C2 (for the advanced class), in the same way as we did for the BNE corpus. This corpus amounts to 154 documents and 23,672 tokens, which conforms a rather small corpus, but we decided to add it to the BNE corpus for achieving more variation in terms of text genre.

3.2 Production Corpus

As our representative of the productive knowledge, we used the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013)[4]. This corpus is divided according to the CEFR (Verhelst et al., 2009), and contains a huge amount of documents written by learners, totaling more than 500 thousand documents with more than 33 million tokens. They were written by 83,385 learners from 137 nationalities, and each document has an evaluation score ranging from 0 to 100 and an associated topic (e.g. *introducing yourself by email*). The data from the corpus is organized following the CEFR classification, so that we just transposed the original division to our own three categories: beginner, intermediate, and advanced.

3.3 Corpus Annotation

Both the reception and the production corpora were automatically annotated with pedagogically relevant grammatical information. For this type of annotation, we used the features described in the SMILLE system (Zilio and Fairon, 2017; Zilio et al., 2017a; Zilio et al., 2017b), which can recognize 107 different grammatical (both morphosyntactic and syntactic) structures in English texts that are relevant for the learning of English as second language[5]. These grammatical structures are ranked according to the CEFR levels, from level A1 to C1, following the pedagogical organization of Altissia's English course. Large part of the information annotated by the SMILLE system is different from a simple parsing annotation, since it represents a combination of parser information with manually created rules that focus on a pedagogical approach to language learning. The annotation was done automatically, and previous work has shown that the overall precision of the system lies around 90.1% for syntactic structures in

[3] www.altissia.com.

[4] We used the EF-CAMDAT version 1 (https://corpus.mml.cam.ac.uk/efcamdat1/) in this paper, because version 2 (https://corpus.mml.cam.ac.uk/efcamdat2/public_html/) was not yet available at the beginning of the research.

[5] For a more complete description of these grammatical structures, please refer to Zilio et al. (2018). The annotated vectors of the EF-CAMDAT corpus can be downloaded from http://cental.uclouvain.be/resources/smalla_smille/sgate/.

the EF-CAMDAT corpus (Zilio et al., 2018) (and precision for morphologic structures is based on the Stanford Parser (Manning et al., 2014) performance)[6].

After the annotation process, each document in both corpora was represented as a vector with the normalized frequency of each of the annotated structures. The normalization of each of the syntactic features was done by dividing their absolute frequency in the document by the number of sentences in it. For the morphosyntactic features, the normalization was carried out by dividing the absolute frequency of the feature in the document by the number of tokens in it.

4 Grammar Profile: Production vs. Reception

The methodology that is generally applied to profiling consists in the use of a training corpus with annotated documents, which are then converted to vectors of features, and methods of classification based on machine learning. The creation of vectors normally takes into account the stylometry of the document. As an example of profiling, Argamon et al. (2009) employ a taxonomy of a specific grammatical class, combining it with other metrics for training a classification algorithm.

By using a corpus of input texts (texts that are used as a receptive means for language learning) and a corpus of output texts (texts that are produced by language learners), we could contrast the occurrences of pedagogically relevant grammatical structures in both of them. This enabled us to come up with a language profile oriented to language learning that considers both sides of the writing spectrum.

In this profiling task, while we used a random selection of 250 documents from each of the three classes (beginner, intermediate, and advanced) of the input corpus (totaling 750 documents), in terms of output, we randomly selected the same amount, but considering only those texts for which the evaluation score given by an expert was between 90 and 95. This ensured, at the same time, that our sample was not made up of texts that were above their level, which is a possible explanation for a maximum score, nor did it contained texts that were not so well formed from a linguistic point of view, because this could pose problems for the automatic annotation and, thus, for our profile.

The annotated corpora were analyzed in terms of the Jensen-Shannon divergence that exists in each of the three classes (beginner, intermediate and advanced) for each of the 107 annotated grammatical structures. As explained in Section 3.3, for a more balanced analysis, the structures were normalized in relation to the number of sentences (in the case of syntactic structures) and tokens (for morphosyntactic structures), and then, for applying the divergence score between both corpora, they were averaged in relation to the number of documents in each corpus. In this process, any structure with a relative frequency equal to zero, whether in the reception or the production corpus, was discarded. The divergence was calculated using ten different random samples for ensuring a better confidence in the data, and the results from the ten samples were afterwards averaged.

After the calculation of the divergence of each of the non-discarded grammatical structures, we had to select a cut-off point for making up our language profile. Since the divergence provides scores, but not a clear cut-off point, we deliberated that the structures classified at the upper quartile regarding the divergence distribution were a good choice to form our language profile. By the end of this process, we identified structures that could be considered problematic in terms of writing for a language learner in each of the three levels. However, the divergence score gave us only the magnitude of the divergence, but it did not account for where the divergence exactly lies, and, for our purposes, we deemed important to know the direction of the divergence, i.e., to know whether a structure is more prominent in the reception or the production. So we went further on the task and used a voting system based on normalized frequency to see which of the corpora presented more of a given structure. This voting system was used to attribute a polarity to the divergence score, representing cases of overproduction (positive scores), which are probably a sign of easier structures for the learners, but that are not that important in a more naturally written text, or cases of underproduction (negative scores), which are an indicator of structures that are more frequent in the input, but the learners may have some difficulties in producing them. For attributing a polarity to the grammatical structure, at least seven out of the ten random samples of the

[6]Further evaluation of more specific structures in different corpora can be found in Zilio et al. (2017a) and Zilio et al. (2017b).

output corpus had to point to overuse or underuse, if there was no side with seven votes or more, we filtered the grammatical structure out of the results, because it represented a case in which the use of the grammatical structures fluctuated too much in the samples.

Many of the highly divergent structures (upper quartile) were similar among the classes, but some of them showed an overuse in some levels and an underuse in others. Table 1 presents the structures that were divergent from input to output along all the levels, and the over- or underuse is expressed by plus or minus symbols, respectively. In the table, lines 1 to 9 represent structures that are divergent in all three classes, from beginner to advanced; lines 10 to 12 show structures that were divergent both in the beginner and intermediate classes; then lines 13 to 22 present structures that were divergent both in the beginner and advanced classes; and, finally, the remaining lines display structures that were divergent in only one of the classes (lines 23 to 26 for beginner, lines 27 and 28 for intermediate, and lines 29 to 34 for advanced). As we can see, the grammatical structures that present divergence go from more coarse-grained syntactic phenomena, such as passive voice and imperative, to more fine-grained ones, like the verb "to be" in different tenses and use of "would" to express hypothesis.

Before going to the other experiments, there are some interesting information presented by our profile that we would like to discuss. For instance: it seems that learners of English tend to overuse the present tense, while neglecting the past tense in relation to the texts that are considered fit for reception. There is also a constant trend to neglect the use of genitive, the present perfect tense and the use of direct complements. These structures that tend to be neglected throughout the levels may be clues for grammatical structures that pose a general problem for language learners. The divergence on the use of direct complements may be also a clue for the existence of a language use with emphasis on intransitive verbs. All these indications present a good basis for further linguistic investigation that would need to be directly assessed in the texts.

4.1 Applying the Grammar Profile

The grammar profile that we generated is intended to describe the behavior of grammatical structures in texts that are either input or output of a language learning task. So, we used it as a basis for other experiments that are important in helping a language learner to improve their writing skills but also that serve as an extrinsic evaluation of the profile.

For these experiments, that we describe and discuss in Sections 5 and 6, we used the grammatical structures that were present in the profile, which amounted to 26 for the beginner class, 14 for the intermediate class and 25 for the advanced class, in terms of their normalized frequency and of their significant difference in relation to the relative frequencies that were learnt in the samples used for generating the profile.

Following this principle, each of the documents in all of the following experiments were rendered as a vector with two scores per grammatical structure in the profile, according to the class to which it belongs: the first score is the normalized frequency observed in the document divided by the normalized frequency from the profile for the reception corpus; and the second score is the normalized frequency observed in the document divided by the normalized frequency from the profile for the production corpus. For both those scores, if there is no significant difference between the normalized frequencies, the score was set to 1.

Since a well-formed text is composed not only of grammar, but also of a series of linguistic information, including some that are not yet computationally verifiable (e.g., certain pragmatical and discursive aspects), we designed the experiments in the next sections in a way that they include not only our profile based on grammar, but also vocabulary and readability measures. By doing so, we are adding to our profile features that are well described in the literature and that may compensate for some of the extra linguistic information that is lacking in the profile. Even so, in each of the experiments, we also tested all of the feature sets separately. This allowed us to observe the possible contribution of others measures to the language profile that we propose here.

For the vocabulary features, we used as basis a frequency list that was extracted from the British National Corpus (BNC) (Aston and Burnard, 1998) and divided in five pedagogically distributed ranges

Table 1: Grammatical Structures: cases of overuse and underuse contrasting the productive with the receptive corpus

		Beginner	**Intermediate**	**Advanced**
1	*genitive*	-	-	-
2	*advanced modal verbs*	-	+	+
3	*past simple*	-	-	-
4	*present perfect*	-	-	-
5	*present simple*	+	+	+
6	*verb "to be" in the present simple*	+	+	+
7	*infinitive with "to" after a verb*	-	+	+
8	*number of direct complements*	-	-	-
9	*present participle*	+	+	-
10	*imperative*	+	+	
11	*number of main verbs*	-	-	
12	*number of subjects*	-	-	
13	*simple modal verbs*	-		+
14	*modal verbs with ellipsed main verb*	-		+
15	*passive voice*	-		-
16	*relative clauses*	-		-
17	*future with "will"*	-		+
18	*verb "to be" in the past simple*	-		-
19	*infinitive with "to" after an adjective*	-		+
20	*connectives of time*	-		+
21	*connectives of reason and result*	+		+
22	*connectives of purpose*	-		-
23	*use of articles*	-		
24	*use of plurals*	-		
25	*use of numbers*	+		
26	*present continuous*	+		
27	*use of present and present perfect after time connectives*		+	
28	*connectives of condition*		+	
29	*use of would to express hypothesis*			+
30	*prepositional verbs*			+
31	*infinitive with "to" after a noun*			+
32	*gerund after prepositions*			-
33	*connectives of alternative*			-
34	*connectives of example and explanation*			+

(Martinez, 2011; Cobb, 2013). Using this list we calculated the frequency of words from each range in each of the documents from our corpora. As a means of normalizing the frequency, we divided the sum of the frequency of words from each BNC range in the document by the total number of words from all ranges in the same document. Finally, for the readability measures, we used the Dale-Chall Score (Dale and Chall, 1948), the Flesch-Kincaid measure (Kincaid et al., 1975), and the Gunning Fog Index (Gunning, 1952), which are commonly used in the literature.

5 Text Classification: Production Corpus

Having a language profile at hand that emphasizes the divergence between reception and production in language learning, we used it to assess how good it is for classifying texts produced by language learners in the respective levels. This is an important task in the goal of helping learners to write texts that match their level, since the first step in evaluating a text is to attribute it to the correct level, so that it is possible

to further analyze its weaknesses and strengths.

For this test, we randomly selected documents from the production corpus that were not used to generate the divergences. For these random selection of documents, we used certain constraints, so that the documents must have been evaluated above 90%, so as to match the criteria employed by Wilkens et al. (2018), who use the same corpus, so that we would have a comparable basis. The corpus was divided in beginner, intermediate, and advanced classes, as explained in Section 3.2. The full sample contained 15,068 documents (6 thousand for beginner and intermediate, and 3,068 for advanced[7]). For the classification task, we used the Random Forest algorithm (Breiman, 2001) with a ten-fold cross-validation.

The results of this experiment are shown in Table 2 in terms of F-measure and standard deviation. For the beginner level, our best F-measure was 0.939, decreasing to 0.858 in the intermediate level and further to 0.822 for the advanced level. By averaging the three levels, we got a non-weighted result of 0.872 (0.882 if weighted). The best average result was achieved by mixing all three measures, and, with exception of the advanced level, which fared a bit better with only vocabulary features, the other levels also had better results using the combined approach. We can also see that the worst results were produced by readability measures, which scored consistently below the other features. Our best results in this experiment are consistent with the state of the art for document classification in levels, as can be seen in the description by Wilkens et al. (2018).

Table 2: F-measure and Standard Deviation for the Level Classification in the Production Corpus

	Profile	**Vocabulary**	**Readability**	**All**
Beginner	0.922 (0.024)	0.893 (0.020)	0.775 (0.030)	**0.939** (0.022)
Intermediate	0.819 (0.000)	0.811 (0.000)	0.589 (0.000)	**0.858** (0.000)
Advanced	0.747 (0.000)	**0.822** (0.000)	0.464 (0.000)	0.819 (0.000)
Average	0.830 (0.009)	0.842 (0.010)	0.609 (0.014)	**0.872** (0.009)

Values in bold reflect the best scores in each level and the average best score. All scores were significantly different from the best scores with a confidence of 99%.

5.1 Text Scoring

This experiment was conducted to observe if our language profile was good for evaluating texts produced by language learners in terms of score (the actual score attributed to a text by an evaluator regarding its well-formedness in relation to the given task). To do so, we randomly selected 70 texts from the production corpus per each ten score points[8] as input for the Random Forest algorithm in a three-fold cross-validation. As such, for each document received as input, the algorithm produced a score from zero to a hundred, in the same fashion as an expert would, for each of the EFCAMDAT documents in the sample. After that, we calculated the correlation of the scores produced by the algorithm with the actual scores that were given by the experts that evaluated the Cambridge Exams.

Using a total of 1,982 documents (734 documents for beginner, 700 for intermediate and 548 for advanced), we ran a three-fold cross-validation. The results in terms of Pearson's correlation and standard deviation are presented in Table 3. Again, as we saw for the classification experiment, the combined features yielded the best results for the beginner and intermediate levels, while the vocabulary alone was better to predict the scores of the advanced level, although this result was not significantly different from the one achieved by the combined features. Looking at the table and disconsidering the readability alone, which was a catastrophe, the correlations, especially for the advanced level, were not so bad, ranging from 0.509 to 0.685. However, by looking at the root mean squared error (RMSE), we see values ranging from 30.42 (advanced level) to 32.14 (beginner level), with a stop at 31.24 (intermediate

[7]This is the whole of EFCAMDAT for C1 and C2 combined for scores above 90% after excluding those that were used for calculating the divergence scores.

[8]We divided the whole corpus in eleven truncated ranges: from 0 to 9, 10 to 19 etc., up to 90-99, and 100 as a separate range, for each of the three classes (beginner, intermediate, and advanced). The EFCAMDAT contains lots of documents evaluated with higher notes, but not so many on the lower side. For some of the lower score ranges, there was not 70 documents, so we used all those available. None of these documents were present in the production corpus that was used for developing the language profile.

level). These values, which were not significantly different between each other in the levels for all setups, are far from good.

These results indicate that other factors need to be taken into consideration for the evaluation of a learner's text in terms of score, not only grammatical and/or vocabulary features, let alone readability measures. It is also important to consider that, for the scores, there is always a strong subjective impact based on the expert's opinion, which is beyond our actual capacity of modeling.

Table 3: Correlation and Standard Deviation for the Scoring Experiment

	Profile	**Vocabulary**	**Readability**	**All**
Beginner	0.457 (0.044)	0.445 (0.042)	0.253 (0.054)	**0.509** (0.034)
Intermediate	0.459 (0.044)	*0.512 (0.034)	0.242 (0.056)	**0.524** (0.037)
Advanced	0.624 (0.038)	**0.685** (0.047)	0.364 (0.055)	*0.674 (0.033)

Values in bold reflect the best scores in each level and the average best score. The values with a star (*) were not significantly different from the best scores with a confidence of 99%.

6 Text Classification: Reception Corpus

This last experiment was designed to test the capacity of our profile to classify texts that are used as input for language learning activities. Although this task does not directly relate to helping a learner in a writing activity, it does serve as an extrinsic test for evaluating the quality of our language profile, while also being an important task in language learning in general.

This experiment followed the exact methodology described in Section 5 for the production corpus, with the obvious difference that here we used the reception corpus as basis and, as such, we had a quite drastic reduction in the number of documents used for the classification, especially because we used only those documents that were not used for generating our language profile. So, to improve the corpus size, we sorted 10 random document samples from those that were not used in the profile. This process yielded a corpus of 1,892 documents[9]. We then used a ten-fold cross-validation with the Random Forest algorithm. Table 4 shows the results for this experiment.

Table 4: F-measure and Standard Deviation for the Level Classification in the Reception Corpus

	Profile	**Vocabulary**	**Readability**	**All**
Beginner	0.891 (0.053)	*0.903 (0.042)	0.878 (0.049)	**0.922** (0.039)
Intermediate	0.861 (0.000)	0.916 (0.000)	0.853 (0.000)	**0.921** (0.000)
Advanced	0.914 (0.000)	**0.984** (0.000)	0.932 (0.000)	0.982 (0.000)
Average	0.889 (0.033)	*0.934 (0.027)	0.888 (0.031)	**0.942** (0.026)

Values in bold reflect the best scores in each level and the average best score. The values with a star (*) were not significantly different from the best scores with a confidence of 99%.

The results of this experiment were not much different from the one that we carried out with the production corpus. We saw a general improvement of the scores, achieving a non-weighted average F-measure of 0.942 (0.940, if weighted). But here the vocabulary features excelled, presenting an average result similar to the combined model, and significantly exceeding the combined model for the advanced class. The language profile alone fared less well for the classification of reception texts, but still beat the readability features for the beginner and intermediate classes.

7 Conclusion

This study aimed at developing a language profile for second language acquisition that considers both productive and receptive knowledge, so that the information from the profile could be used as a helping tool for the learners to write a more naturally sounding text. We also hypothesized that, by adding

[9]Many of the documents in the corpus are duplicated due to the ten random samples that were selected.

grammatical information to already tested and proven vocabulary and readability features, we would achieve better results in terms of SLA-related tasks.

For achieving these goals, we used two corpora, one representing the productive knowledge and the other the receptive knowledge. We annotated both corpora with pedagogically relevant grammatical information and calculated a divergence score for each of the grammatical structures to find out which ones were different in the comparison of both corpora.

The results of the divergence score were further improved by an analysis of overuse or underuse of structures in the production corpus in relation to the production corpus. This improvement to the divergence score resulted in a grammatical profile that shows where and how the written texts from language learners differ from reception texts. The structures shown in Table 1 also present clues for grammatical structures that are possibly too easy or for those that remain a difficulty throughout the learning process.

By conceiving experiments for the classification of production and reception texts, we could observe that our profile can very well model the written texts, achieving results that are comparable to those of true and tested vocabulary and readability features. Regarding our hypothesis, the addition of grammar to lexical information did yield the best results in the production corpus, but the lexical information alone was enough to achieve the same result as the combined approach in the reception corpus.

It is important to emphasize that our results for the classification experiment in terms of the reception corpus are competitive with the state of the art. For instance, Xia et al. (2016), by means of a profound analysis of parameters, achieved an F-measure of 0.845 using readability features and a lexical distribution similar to our vocabulary features. The same is true for the classification experiment regarding the production corpus, since we achieved similar results to those presented by Wilkens et al. (2018).

In terms of the scoring experiment, we observed that scoring is a very challenging task and that there is still need for further study before we can get good results. Even so, we did get some encouraging results, reaching almost 0.7 correlation for the advanced class in a task that is known to present high discordance among evaluators.

As future work, we intend to increase the size of the reception corpus by including texts from didactic material that is used in second language classes. We are also going to develop a system for supporting the writing of texts that takes as basis our grammar profile, by detecting deviating behavior of grammatical structures in texts produced by the user.

Acknowledgements

The authors would like to thank the Walloon Region (Projects BEWARE n. 1510637 and 1610378) for support, and Altissia International for research collaboration.

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Guy Aston and Lou Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Capstone.

Sean Banville. 2009. Breaking news english. *Teaching english as Second or Foreign Language*, 13(1).

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Thomas Cobb. 2013. Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, pages 79–108.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

May Fan. 2000. How big is the gap and how to narrow it: an investigation into the active and passive vocabulary knowledge of l2 learners. *An investigation into the active*.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 31. IEEE.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*.

Robert Gunning. 1952. The technique of clear writing.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

Batia Laufer. 1998. The development of passive and active vocabulary in a second language: same or different? *Applied linguistics*, 19(2):255–271.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Ron Martinez. 2011. *The development of a corpus-informed list of formulaic sequences for language pedagogy*. Ph.D. thesis.

Paul Meara. 1990. A note on passive vocabulary. *Interlanguage studies bulletin (Utrecht)*, 6(2):150–154.

BQ Morgan and Lydia M Oberdeck. 1930. Active and passive vocabulary. *Studies in modern language teaching*, pages 213–221.

Michael P Oakes and Malcolm Farrow. 2007. Use of the chi-squared test to examine vocabulary differences in english language corpora representing seven different countries. *Literary and linguistic computing*, 22(1):85–99.

Michael P Oakes. 2008. Statistical measures for corpus profiling. In *Proceedings of the Open University Workshop on Corpus Profiling, London, UK (October 2008)*.

Ferdinand Österreicher and Igor Vajda. 2003. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653.

Virginie Pignot-Shahov. 2012. Measuring l2 receptive and productive vocabulary knowledge. *Language Studies Working Papers*, 4(1):37–45.

Norbert Schmitt. 2008. Instructed second language vocabulary learning. *Language teaching research*, 12(3):329–363.

Norman Verhelst, Piet Van Avermaet, Sauli Takala, Neus Figueras, and Brian North. 2009. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Rodrigo Wilkens, Leonardo Zilio, and Cédrick Fairon. 2018. Sw4all: a cefr classified and aligned corpus for language learning. In *Proceedings of the 11th Language Resources and Evaluation Conference*.

Menglin Xia, Ekaterina Kochmar, and E Briscoe. 2016. Text readability assessment for second language learners.

Leonardo Zilio and Cédrick Fairon. 2017. Adaptive system for language learning. In *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on*, pages 47–49. IEEE.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2017a. Enhancing grammatical structures in web-based texts. In *Proceedings of the 25th EUROCALL*, pages 839–846. Accepted.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2017b. Using nlp for enhancing second language acquisition. In *Proceedings of Recent Advances in Natural Language Processing*, pages 839–846.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2018. An sla corpus annotated with pedagogically relevant grammatical structures. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may. European Language Resources Association (ELRA).