Argentinian Journal of Applied Linguistics
Vol. 7, No. 2, November 2019, 34-51

Vocabulary selection for didactic purposes: report on a machine learning approach

Patrick Goethals* - Ghent University, Belgium
Arda Tezcan - Ghent University, Belgium
Jasper Degraeuwe - Ghent University, Belgium

(Received 14/12/18; final version received 31/08/19)

ABSTRACT

This paper aims to make an innovating contribution to the field of technology-enhanced vocabulary learning. We report on a machine learning experiment that supports vocabulary item selection for didactic purposes. We tested two machine-learning algorithms to predict the difficulty level of lexical items as reported by intermediate-advanced learners of Spanish as a foreign language and analyzed the predictive power of various features on this task. This methodology can be especially useful in data-driven autonomous learning contexts.
This paper aims to make an innovating contribution to the field of technology-enhanced vocabulary learning. We report on a machine learning experiment that supports vocabulary item selection for didactic purposes. We tested two machine-learning algorithms to predict the difficulty level of lexical items as reported by intermediate-advanced learners of Spanish as a foreign language and analyzed the predictive power of various features on this task. This methodology can be especially useful in data-driven autonomous learning contexts. It makes it possible to create adaptive environments that select the most appropriate target items for different types of vocabulary learning activities. We will describe the empirical results of the experiments, and will also show how the methodology is integrated in an on-line learning environment.
Keywords: vocabulary learning; vocabulary selection; Spanish; machine learning

RESUMEN

Este trabajo pretende hacer una contribución innovadora a la enseñanza del vocabulario asistida por la tecnología. Describimos los resultados de un experimento de aprendizaje automático que ayuda a seleccionar elementos de vocabulario con fines didácticos. Se analizan dos algoritmos para predecir el nivel de dificultad de los elementos léxicos, definido por estudiantes de español como lengua extranjera de nivel intermedio-avanzado y se ha analizado el poder predictivo de diferentes variables. La metodología puede ser especialmente útil en contextos de aprendizaje autónomo basado en datos. Permite crear entornos interactivos que seleccionan los elementos más apropiados

* Corresponding author, e-mail: patrick.goethals@ugent.be

para diferentes actividades de aprendizaje de vocabulario. Describiremos los resultados empíricos de los experimentos, así como la manera en que se integra la metodología en un entorno de aprendizaje en línea.
Palabras clave: aprendizaje de vocabulario; selección de vocabulario; español; aprendizaje automático

WITH THIS PAPER we will contribute to one of the main discussions in the field of vocabulary learning (Groot, 2000; Nation, 2016), namely the debate on identifying the most appropriate empirical principles that can guide the selection and difficulty grading of vocabulary items. In particular, we aim to show that a machine learning approach can improve and facilitate this selection and grading process, especially in order to calibrate the different parameters that could have an impact on the difficulty level of a lexical item. Moreover, we will show that the automatization of this selection process opens new perspectives for autonomous and customized learning. Although we work with one target language (Spanish) and one target group of (Dutch-speaking) learners, the results are useful for the broader field of technology-enhanced vocabulary learning and teaching.

Importantly, the research that will be presented is embedded in a didactic environment, by which we mean that the results are directly implemented in a learning tool. In the sections that follow, we will describe this environment (1), present a brief overview of current insights on vocabulary selection and difficulty grading (2), describe the empirical methodology that was used in the experiments (3), present the results of the machine learning experiments (4) and, finally, evaluate the didactic implications of the research (5).

The SCAP project

As mentioned before, this paper is part of a broader Research & Development project (Goethals, 2018). The aim of SCAP is to develop annotated corpora and algorithms that support data-driven and corpus-based vocabulary learning processes (Boulton, 2017; Little, 2007) by combining techniques and insights from Natural Language Processing, Corpus Analysis and Second Language Learning. These algorithms, the corpora and the didactic outcomes are integrated in a web-based learning interface (https://scap.ugent.be).

It should be noted that the SCAP vocabulary learning application is not aimed at beginner students but, on the contrary, is designed to fulfill the pedagogical needs of high-intermediate and advanced learners of Spanish (B2+) who wish to develop their vocabulary knowledge in specific domains. It may be challenging to satisfy these needs in a classroom setting because the interests of advanced learners may vary considerably, and it is crucial to motivate the students by selecting semantic domains that belong to their interest domain (Salazar García, 2004). Therefore, the application is designed to parse a corpus representing a semantic domain chosen by the student, and to generate on this basis learning materials such as vocabulary lists, glossaries, cloze exercises or reading text selection (see also

Section 5 for some examples). An appropriate selection of the vocabulary items is crucial, both to guarantee that the vocabulary items are indeed specific or at least relevant for the semantic domain, and that they are maximally adapted to the proficiency level of the learners. In this paper we will mainly focus on the proficiency level, and less on the semantic domain specificity. As will become clear from the literature review below, one of the specific characteristics of SCAP is that we develop a vocabulary selection method for advanced levels, whereas most proposals have focused on defining the "first" ranges of vocabulary items, varying between 5000 and 9000 lexical items. The focus on advanced vocabulary learning is challenging, because there is no clear "zero point", and because after the first threshold of, for example, 5000 words, we enter into a very diffuse field before reaching the 20.000-30.000 words used by higher-educated native speakers, let alone the 100.000 words of a general dictionary (Santos Palmou, 2016). Given the almost infinite number of possible lexical items that may occur in specialized fields, it is too time-consuming to manually check these vocabulary selections, at least if the aim is to cover a broad range of semantic domains. Therefore, it is an important challenge to automatize this process.

Vocabulary selection

As can be inferred from previous literature reviews, vocabulary selection and grading are considered crucial but complex steps in the design of didactic materials (Bartol Hernández, 2010; Nation, 2016; Vincze and Alonso Ramos, 2015). It has become common practice to complement or substitute introspective methods by empirical and mainly corpus-based methodologies. Within a corpus-based methodology, the most obvious parameter is of course frequency, the assumption being that the most frequent words in a corpus are also the most interesting or useful ones (see Davies, 2005 and 2006 for Spanish). Yet, many authors have argued that raw corpus frequencies should be handled carefully, and corrected, for example, by:

- improving the representativeness of the corpus, e.g. by building a corpus that includes a sufficiently wide variety of text types (Davies, 2006), or by validating the representativeness of different corpora (Duchon et al., 2013;
- complementing the overall frequencies with data on the distribution or dispersion of words throughout the corpus (Davies, 2006; Gries, 2008; Nation, 2016);
- taking into account cognate effects between words in the target language and the mother tongue of the students (Izquierdo Gil, 2005);
- critically evaluating the outcome of the empirical selection procedures by taking into account the intuition of experienced teachers or didactic authors (Instituto Cervantes).

We take these insights as a starting point to explore the question that inevitably follows the identification of possibly relevant factors, namely how these factors can be calibrated and

combined into one single selection procedure. As will be explained in the next section, we propose an experimental machine learning approach, in which we will (a) create a gold standard consisting of students' evaluations of the difficulty level of a collection of lexical items (dependent variable), (b) gather data representing frequency in different corpora, dispersion between corpora, cognateness and independently assigned difficulty levels (independent variables), and, finally, (c) evaluate the prediction performance of different machine learning systems that are trained on these data and analyze the predictive power of different types of features (independent variables) using a feature selection method.

Data

Corpus and target item selection

For this case study we work with a 273K words corpus on a specific business communication domain, namely CEO and CFO presentations at stakeholder meetings of Spanish companies. The target audience for the didactic application could be, for example, interpreters preparing themselves to interpret these presentations, multilingual employees of financial institutions attending these meetings, trainees in a course of Spanish for specific purposes, or teachers preparing didactic materials for these trainees.

The corpus was part-of-speech tagged and lemmatized with the SCAP pipeline (Goethals et al., 2017), and lemmalists were generated for the noun, verb and adjective part-of-speech categories. From these lists (4900 lemmas in total) we extracted a list of possibly relevant target items for advanced learners (B2+) by applying the following criteria:

- we removed all lemmas included in the thematic word list *Thematischer Grund- und Aufbauwortschatz Spanisch* (Navarro and Navarro, 1996/2010, also adapted for Dutch-speaking ELE students), comprising +/- 5000 lemmas. The reason for choosing this particular word list is that it is used in the program followed by the participants in the experiment: it could reasonably be expected that these words would be overwhelmingly judged as "known" by the participants, which makes them uninteresting candidates for the selection and ordering experiment. Moreover, as was already said, we are not interested in delimiting the "first" ranges of vocabulary to be learned, but rather in organizing everything that comes behind the basic-intermediate threshold;
- we only kept those items that were significantly more frequent in this corpus than in an ad hoc created reference corpus representing non-business discourse (concretely, tourism leaflets; for more details on Keyness calculus in SCAP, see Goethals, 2018);
- English loan words such as *online* or *web* were removed;
- when the lists contained two closely related lemmas belonging to the same word family (e.g. *concentrar – concentración*) we only kept the base form (in this case the infinitive).

The result was a total number of 531 items (230 nouns, 119 verbs, 182 adjectives; See Table 2 in Section 3.2) that serve as the target items of the experiment.

Dependent variable: students' difficulty judgments

In order to define the dependent variable of "difficulty level", a group of students[1] with estimated vocabulary proficiency levels ranging from B2 to C1 was asked to choose the most appropriate statement below for each target item.

A	I understand this word, and I would use it spontaneously
B	I understand this word, but I would not use it spontaneously
C	I do not understand this word

Although this still represents a simplification of what it means to "know" a word (Nation, 2016), the wording of the initial question invites the students to distinguish between comprehension- and production-oriented knowledge of a word. On the basis of these judgments we defined four categories, with two poles consisting of words that are "known" and "new", and two intermediate categories:

label	**criterion**	**functional description**
A	more than 2/3 of the students chose A	these items appear to be sufficiently known by most B2+ students and it does not seem necessary to include them in explicit vocabulary learning activities
A-B	less than 2/3 of the students chose A and the sum of A+B is higher than the sum of B+C	these items are sufficiently understood and can be used directly in production-oriented activities (e.g. cloze sentences, or sentence writing)
B-C	less than 2/3 of the students chose C and the sum of B+C is higher than the sum of A+B	these items seem rather challenging and will be used first in comprehension-oriented activities (e.g. reading contexts, glossaries, recognition tasks) and then in production-oriented activities
C	more than 2/3 of the students chose C	these items seem very challenging for most students, and it may be advisable to use them only in comprehension-oriented tasks

Table 1: Operationalization of the difficulty level assignment

The results of the student survey are summarized in Table 2, with a total number of 219 items at level A, 157 at level A-B, 118 at level B-C and 37 at level C.

	Difficulty level				
	A	**A-B**	**B-C**	**C**	**Total**
NC	97	60	51	22	230
V	57	31	22	9	119
ADJ	65	66	45	6	182
Total	219	157	118	37	

Table 2: Number of target items per POS-category and assigned difficulty judgments

To be complete, it is important to note that in the statistical analysis the A, A-B, B-C and C values are treated as numerical variables ranging from 1 up to 4.

Independent variables

In the experiment, we want to explore the possibility of developing a system that is able to predict this difficulty judgment on the basis of parameters that are more easily accessible than the time-consuming questionnaire methodology that defined the dependent variable. The following variables were generated:

Frequency data

Frequency data were gathered from two non-business related corpora (a 7.5M corpus of youth literature and a 120K corpus of tourism leaflets), that were tagged and lemmatized following the same parameters as the target corpus.

Importantly, the frequency data are not only represented as absolute frequencies but also as ranked frequency groups. It is worth considering this in detail, because, as we will see in the Results Section, this had a major impact on the predictive power of frequency information on the task of predicting difficulty judgments. We used a total of 7 frequency groups: the lowest ranked group contains the items that occur in the target corpus but not in the reference corpus. The second lowest includes all "hapax" items, occurring only once in the reference corpus: we decided to separate these items because, depending on the size and type of corpus, they represent up to 30-40% of the words, which makes the percentile scores of the other items less meaningful. Finally, nouns, adjectives and verbs occurring more than once were subdivided into 5 percentile groups, representing the 0-20%, 20-40%, 40-60%, 60-80% and 80-100% most frequent non-hapax lemmas of the same part-of-speech category. In other words, frequency groups are defined within the same part-of-speech category.

Dispersion between corpora (Keyness)

A Keyness score compares the frequency of the item in the target corpus with its frequency in the reference corpora (%Diff calculus, Gabrielatos and Marchi, 2011, see also Goethals, 2018). One of the most difficult decisions concerning this measure relates to handling the cases where the item does not occur in the reference corpus, since this inevitably implies a division by zero (see Gries, 2008 for a critical review). We decided to assign a score

immediately higher than the highest scores obtained by the other elements. Similar to the frequency data variable, for the Keyness score we also used both absolute numbers (namely the outcome of %Diff) and percentile groups.

Cognate score

A dictionary of Spanish-Dutch translations of the items was created by scraping various Internet sources, including free translation dictionaries and machine translation tools. Then, a "MatchSequence" score was calculated[2], representing the degree of orthographic similarity between the target word and one of its possible translations in the mother tongue of the students (see Table 3 for some examples). The "cognate" feature was defined as 0 or 1, depending on whether the algorithm identified a possible translation with a Matchsequence score higher than 0,66.

ES	**NL**	**MatchSequence**
dividendo	dividend	0,94
diversificación	diversificatie	0,83
integración	integratie	0,76

Table 3. "Cognate" feature. Examples of Matchsequence scores

Graded vocabulary lists

Finally, we included information from the thematic word list *Portavoces* (Buyse et al., 2004[3]). This method is based on corpus data, but enriched by didactically motivated judgments of experienced teachers and didactic authors (see Section 2). It contains a total number of approximately 9000 lemmas. The items were assigned one of the two proficiency levels used in this publication, or a third value if they did not occur in it. We chose to use this reference point because the students who participated in the experiment did not use this method. Another possible reference point would have been the MECR lists published by the Instituto Cervantes, but, since the students had worked with a manual that closely follows the MECR levels, this would possibly have biased the data.

Summarizing this section, the features are listed with their corresponding codes:

freq_abs_ref_1:	frequency, in absolute numbers, in corpus 'youth literature'
freq_group_ref_1:	frequency groups in corpus 'youth literature'
freq_abs_ref_2:	frequency, in absolute numbers, in corpus 'tourism leaflets'
freq_group_ref_2:	frequency groups in corpus 'tourism leaflets
keyness_ref_1:	keyness compared with corpus 'youth literature'
keyness_ref_2:	keyness compared with corpus 'tourism leaflets'
cognate:	cognate score ES-NL
voc_method:	level definition in vocabulary method Portavoces

Analysis

A first exploration: applying ordinal logistic regression

As a first step in the exploration of the data, we applied an ordinal logistic regression in SPSS for every independent variable.

	Model Fitting	Goodness-of-Fit	Pseudo R^2 (Nagelkerke)
freq_abs_ref_1	x	-	9,8%
freq_group_ref_1	x	x	18%
freq_abs_ref_2	x	-	11,5%
freq_group_ref_2	x	x	19%
keyness_ref_1	x	x	7,1%
keyness_ref_2	x	x	7,4%
cognate	x	x	18,7%
voc_method	x	x	21%

Table 4: Main results of the one-factor Ordinal Logistic Regression Analysis. "x" confirms the Model Fitting ($p < 0.05$), and Goodness-of-Fit ($p > 0.05$). Pseudo R^2 score shows the variance in the dependent variable that is explained by this factor.

From these results, some preliminary conclusions can be drawn. First, as could be expected from the literature, both frequency data (especially the two criteria with frequency groups 'freq_group_ref_1' and 'freq_group_ref_2'), cognateness ('cognate') and existing vocabulary gradations ('voc_method') allow to predict a considerable (and perhaps surprisingly comparable) degree of variance in the dependent variable (18-21%). This predictive value is statistically significant ("model fitting"), and the model fits the data sufficiently well ("goodness-of-fit").

Secondly, we see that the predictive power of the frequency data is considerably higher when they are grouped in (manipulated) frequency groups ('freq_group_ref_1' and 'freq_group_ref_2') than when they are treated as absolute frequencies ('freq_abs_ref_1' and 'freq_abs_ref_2'). The latter data have significantly lower Pseudo R^2 scores (9,8% - 11,5% versus 18% - 19%) and they score negatively for the Goodness-of-Fit test. Given these results, we plan to conduct more elaborated statistical analyses in the future, in order to evaluate different models of building frequency rankings, especially for handling items with zero or low frequencies.

Finally, the Keyness data ('keyness_ref_1' and 'keyness_ref_2'), reflecting the specificity of the items for this particular corpus compared with one of the two reference corpora did not perform well. This might be not very surprising since we applied an initial selection of the vocabulary items, removing those items that were clearly not specific or typical. In this sense, it is better to conclude that the Keyness data do not seem to have a clear effect on the further grading of the items once an initial selection has been realized.

Machine Learning experiments

The goal of the machine learning experiments is to train a model that uses the independent variables (features) to predict the dependent variable (difficulty judgments) on unseen data. Prior to training machine learning models, we divided the data into a training set (90%) and a test set (10%), showing a similar distribution with respect to POS-categories and difficulty judgments (see Figure 1 above). We used the same set of features as in the SPSS ordinal regression, with the only difference that we also added the features of the POS-categories. We carried out the experiments with the Python sklearn module, and concretely used two types of machine learning algorithms, namely a linear regression model, which is good at capturing linear relationships between the dependent and independent variables, and a decision tree model, which can capture non-linear relationships. As we don't know the type of relationship between these variables prior to building machine learning systems, the best option is to compare the results of both methods before making conclusions. The models can be evaluated according to different criteria, of which we will use two: Mean Absolute Error (MAE) (as primary evaluation criterion), which is calculated as the average of the absolute differences between the predicted and the "correct" values of the dependent variable (the lower the better), and Pearson correlation coefficient, which measures the correlation between the predicted and real values (the higher the better, in a span of -1 up to +1). As a baseline, we train models that utilize all features that are outlined in the previous section and the POS-features that are described in this section. Table 5 shows the estimation performance of these baseline systems with respect to both evaluation criteria.

	MAE △	Pearson ♿
Linear Regression	0,836	0,246
Decision Tree	0,836	0,209

Table 5. The estimation performance of the models trained using linear regression and decision tree on the test set with respect to MAE and Pearson correlation score.

Even though we define a number of features for predicting difficulty judgments, which we consider relevant for this task, it is not clear if all of these features will be considered useful by the machine learning models we build. For this reason, besides training models that utilize all features, we also apply a feature selection method, namely SFFS (Sequential Forward Floating Selection) (Pudil et al., 1994) to let the machine learning algorithms seek a minimal subset of features that maximise prediction performance. The basic idea behind the SFFS method is that it starts with an empty set of features and successively adds features, provided that this improves the estimation performance. In addition to providing a feature subset, forward feature selection methods allow for analysing the impact of adding individual features on estimation performance at each feature addition step. Moreover, SFFS also performs a feature removal step after each addition step, provided that removing a feature improves the estimation performance. SFFS therefore samples a large number of

feature combinations as feature subsets and has been shown to perform well among the sequential search algorithms (Ferri et al., 1994; Kudo & Sklansky, 2000).

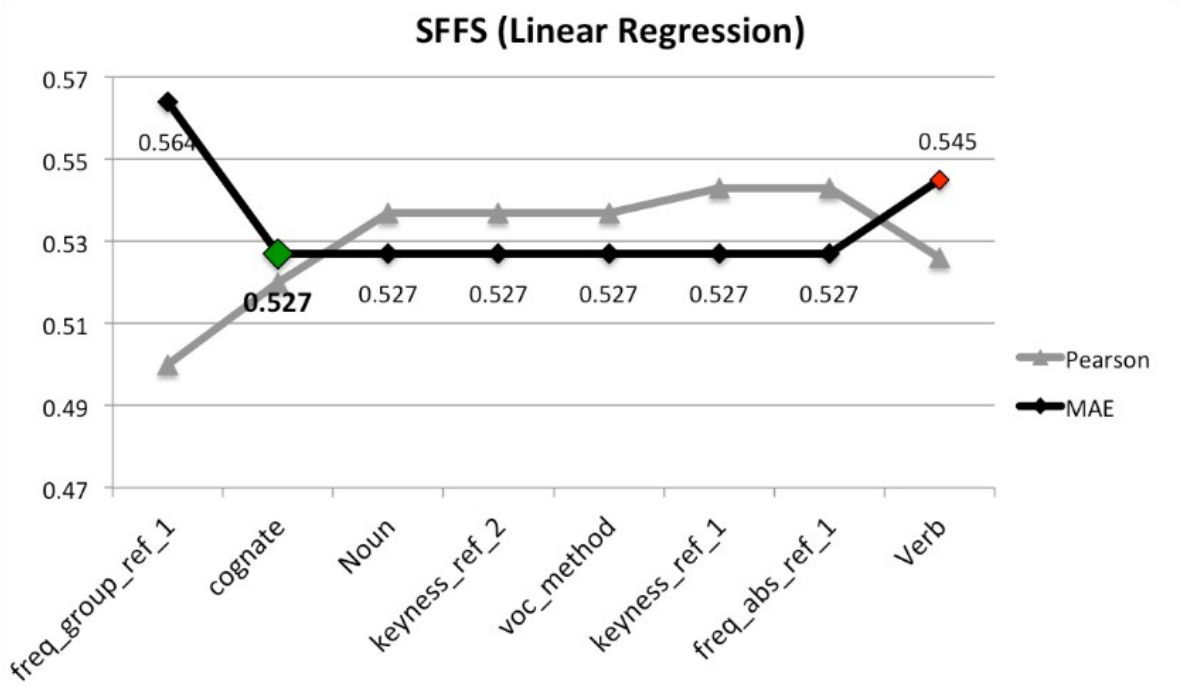


Figure 1. The estimation performance at each SFFS step for linear regression, with respect to MAE and Pearson score. At each SFFS step the given feature is added to the feature subset that contains the features to the left of it[4].

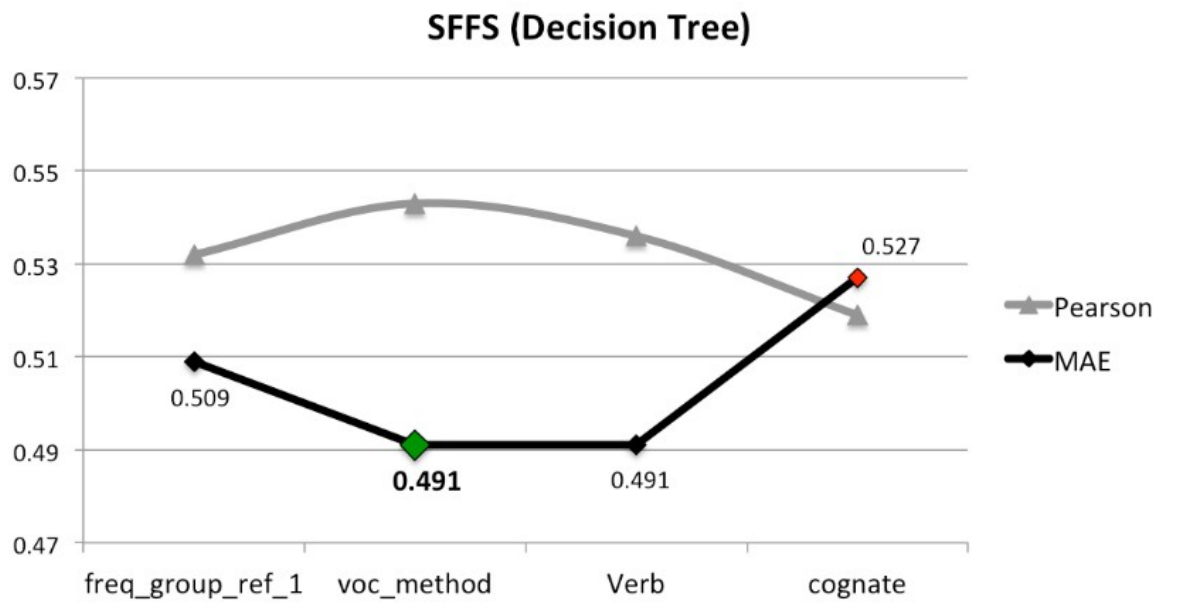


Figure 2. The estimation performance at each SFFS step for decision tree, with respect to MAE and Pearson score. At each SFFS step the given feature is added to the feature subset that contains the features to the left of it.

The first observation we can make from Figures 1 and 2 is that by only using two features, we can build models that outperform the models that use all features and the best performing feature subsets reduce the MAE scores by +/- 40% (0,836 to 0,527 for linear regression and 0,836 to 0,491 for decision tree). In other words, not all features are necessary for predicting

difficulty judgments in this task as combining all features together leads to larger error margins.

The best MAE score for the linear regression model was achieved by the minimal feature subset consisting of 'freq_group_ref_1' (frequency groups in the largest reference corpus of youth literature) and 'cognate' (the existence of a cognate element in the mother tongue of the student) (MAE = 0,527). After these two features, adding other features still points to slight improvements with respect to the Pearson Correlation Coefficient, but not for MAE.

The best MAE and Pearson scores for the decision tree model were obtained by the feature set consisting of 'freq_group_ref_1' (as in linear regression) and 'voc_method' (the assigned level in an independent vocabulary method) (MAE = 0,491). By using only these two features, the decision tree model also outperformed the best linear regression model. This observation can be attributed to non-linear relationships that can be captured by decision tree (and not by linear regression), suggesting that the relationship between the dependent and the independent variables can be explained better by a non-linear relationship in this task. Adding more features did not improve the decision tree model further neither for MAE nor Pearson score.

One interesting observation for the two machine learning algorithms is that they both find 'freq_group_ref_1' very useful. In fact, both algorithms find this feature to be the most useful feature given that it is selected in the first step of the SFFS process for both algorithms. It seems that the two algorithms do not agree on the additional features they find useful. While linear regression uses 'cognate' as a second feature to improve the prediction performance further, decision tree achieves improvements with the feature 'voc_method' instead. Considering the superior prediction performance achieved by decision tree, we consider 'freq_group_ref_1' and 'voc_method' as the best performing feature subset for this task. The errors made by the best system are further analysed in Figure 3.

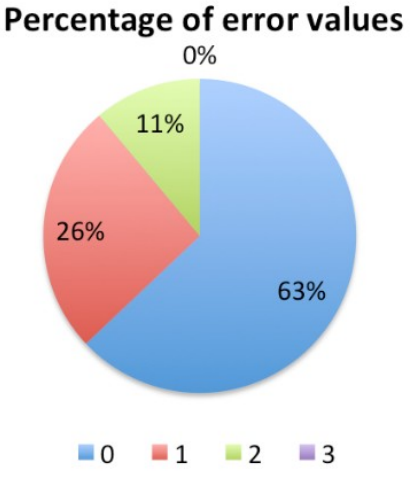


Figure 3. Percentage of the error values between 0 (no error) and 3 (max. error) for the predictions obtained on the test set.

Figure 3 shows us that 63% of all the predictions obtained by the best decision tree model were actually correct. Moreover, the predictions never achieved the highest possible error value of 3, i.e. never predicted difficulty judgment of 1 when the reference value was 4 or the other way around. The model predicted 26% of all the reference values with an error margin of 1 and 11% of all values with an error margin of 2. As a result, this analysis serves as an alternative evaluation method besides MAE and Pearson correlation scores and helps us understand the error profile of the best performing machine learning model we built for this task.

Discussion

This is, by our knowledge, one of the first attempts to use machine learning in the prediction of vocabulary difficulty, especially with respect to difficulty levels beyond the initial or low-intermediate thresholds. Although the dataset is relatively small, the first results look very promising. The best models that were generated have a reasonably limited mean average error, especially if we take into account that the categories are not as clear-cut or mutually exclusive as could be the case in other classification tasks. Moreover, the Pearson Correlation Score shows that there is a clear correlation between the score given by the students and the score given by the algorithm, and this is confirmed by the accuracy score (63%).

Although the aim of this paper was not to discuss on a conceptual level which criteria should guide vocabulary selection, but rather to search for practical solutions for calibrating possible features, we can conclude that the most powerful feature for selecting and grading vocabulary at high-intermediate and advanced levels seems to be the traditional feature of frequency in a general reference corpus (in this case youth literature). This is undoubtedly an important finding, since this feature can be used for all target students, independently of their mother tongue or learning method. On a methodological level, it is important to note, however, that the predictive power of the frequency feature increases considerably when we do not use absolute frequencies but frequency groups. Further research is required to search for the optimal group constructions, but the creation a specific class for hapax items and the fact that we calculate percentile groups within the same POS-category already seem to be powerful hypotheses. However, we would like to add that the power of the frequency feature at high-intermediate and advanced levels does not necessarily imply that it should also be the leading feature at initial or low-intermediate levels. It is very well possible that at these levels it is important to take into account other phenomena such as “lexical availability” (“disponibilidad léxica”, Bartol Hernández 2010) or “imageability” (Duchon et al. 2016).

Apart from the frequency feature, two other features come into the picture: cognateness, which is calculated on an objective basis, and existing vocabulary gradings, which is a mixed feature that also includes a subjective dimension. Both features allow

improving the results based on frequency alone, with the latter outperforming the former. Both features have be adapted to the specific target group of learners: cognateness is calculated according to the mother tongue of the learner, and the feature of existing graded vocabulary lists must take into account which didactic materials were already used by the students.

In addition to the theoretical interest of predicting vocabulary difficulty, we want to emphasize the catalyst effect that the automatization of this process can have for didactic purposes. As was said above, the SCAP project aims to develop not only the theoretical algorithms but also the didactic tools and platforms to bring them into everyday didactic practice. Some of these possibilities have already been integrated in the current version of the platform, and others will be developed in the near future. We will briefly illustrate this with a concrete example. Amongst other functionalities, SCAP allows the user to select a corpus that represents a semantic domain (in this case the corpus of shareholder meetings, ES "juntas de accionistas") and then generates a set of didactic materials, in this case a translation glossary (Figure 4, "glosario breve"). The user can input the lexical items by manually copy-pasting items from the lemma lists that are generated by the tool (Figure 5), but there is also another option, shown in Figure 4, namely that the tool itself makes a selection of the items to be included in the glossary. Currently, there is one predefined option, provisionally called "AFE", which means "avanzado/advanced", "frecuente/frequent" and "específico/specific". This means that the selection is restricted to lemmas that occur relatively frequently in this corpus, are also more frequent in this corpus than in other corpora, and that they are interesting items for "advanced learners". In the current version of SCAP, the "advanced" character of the items is still defined in a very pragmatic way, namely as the list of items that do not occur in the vocabulary method that the students in our institution use, but it is clear that the algorithm that is developed in this paper will allow refining the selection procedure. Figure 6 shows a partial result of this action.

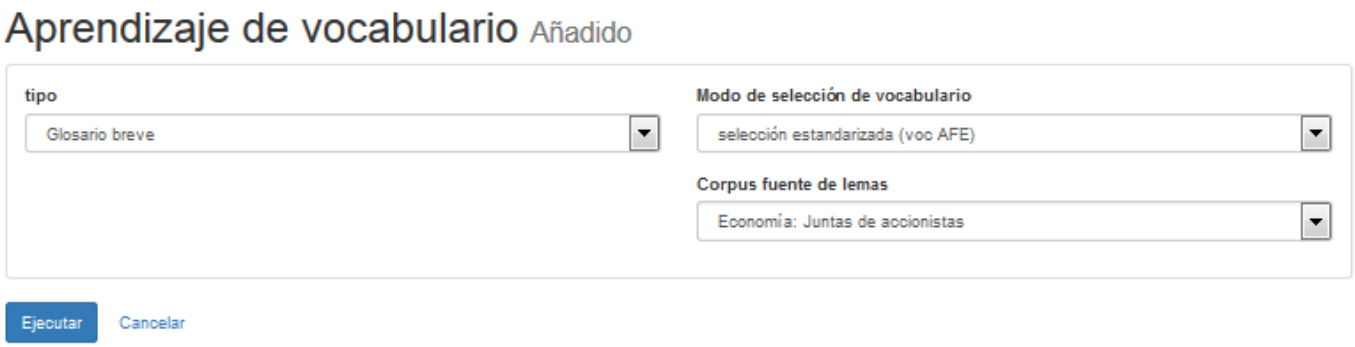


Figure 4: Generating a glossary for the shareholder meeting corpus with automatized vocabulary selection

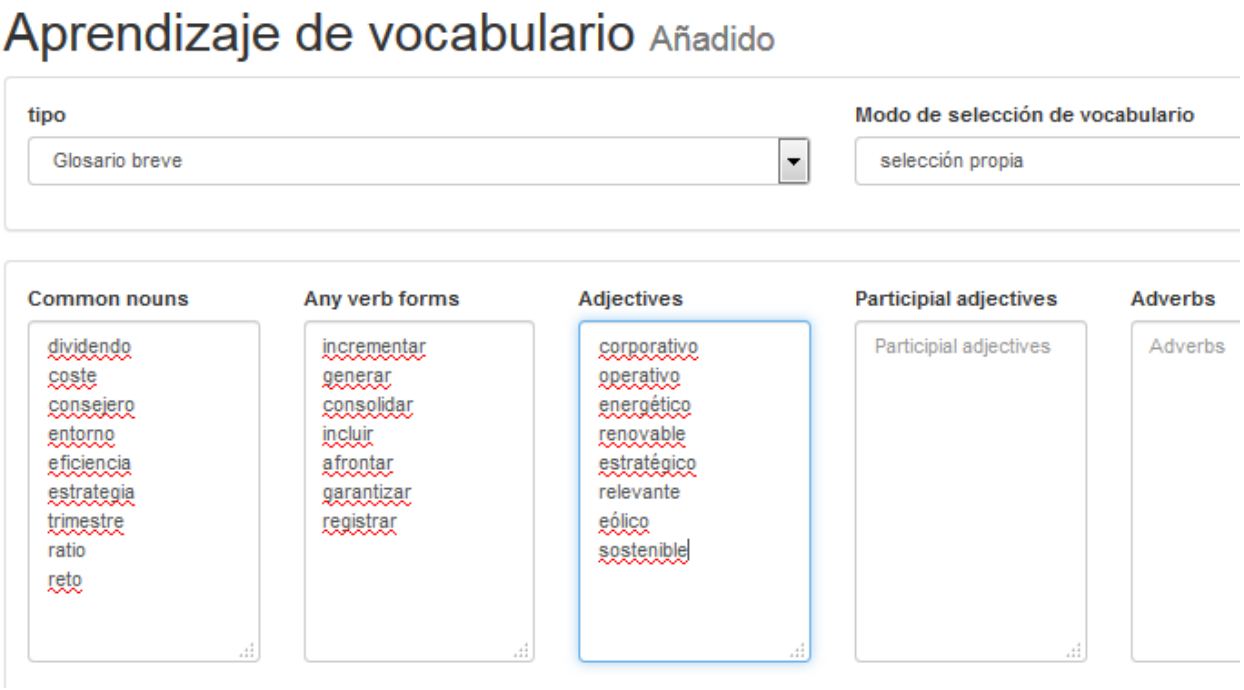


Figure 5: Generating a glossary with manual introduction of vocabulary items

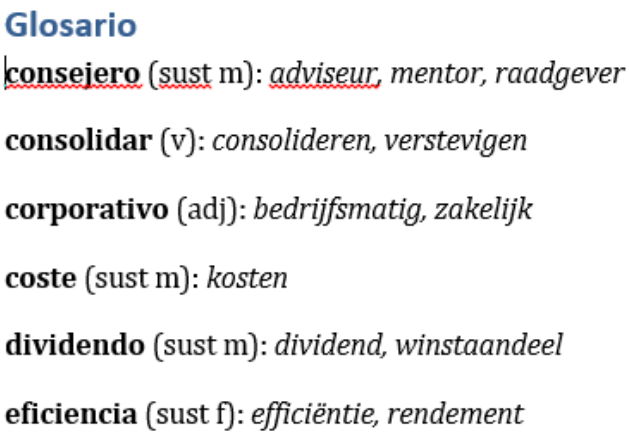
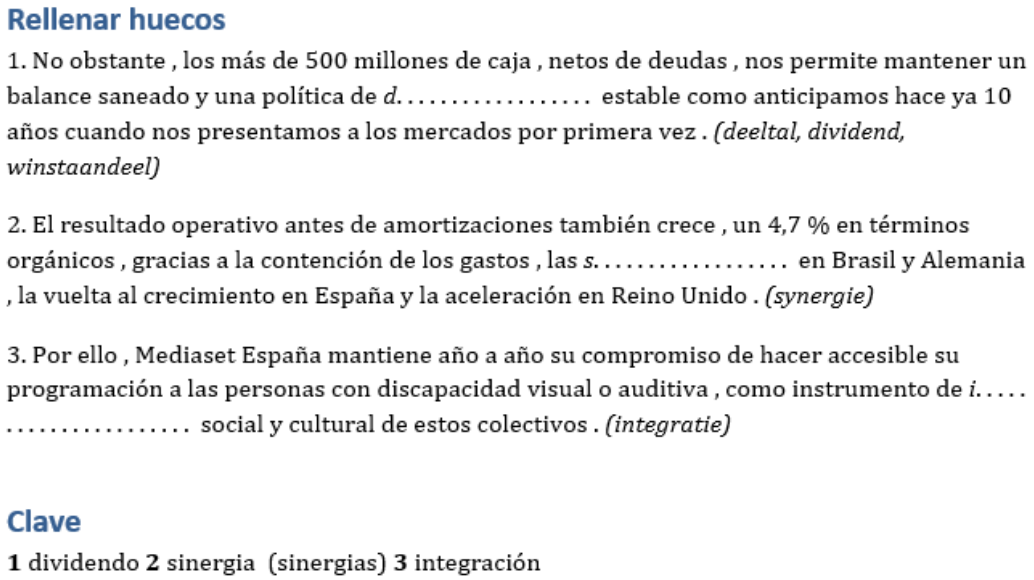


Figure 6: Automatically generated Translation Glossary (Spanish-Dutch).

A similar procedure can be applied for cloze sentences ("rellenar huecos"): in this case the user would find automatically generated fill-in exercises based on the corpus, with some hints such as possible translations from the glossary or the first letter, and the solutions at the end of the document (Figure 7).



Rellenar huecos

1. No obstante , los más de 500 millones de caja , netos de deudas , nos permite mantener un balance saneado y una política de *d.*..... estable como anticipamos hace ya 10 años cuando nos presentamos a los mercados por primera vez . *(deeltal, dividend, winstaandeel)*

2. El resultado operativo antes de amortizaciones también crece , un 4,7 % en términos orgánicos , gracias a la contención de los gastos , las *s.*..... en Brasil y Alemania , la vuelta al crecimiento en España y la aceleración en Reino Unido . *(synergie)*

3. Por ello , Mediaset España mantiene año a año su compromiso de hacer accesible su programación a las personas con discapacidad visual o auditiva , como instrumento de *i.*..... social y cultural de estos colectivos . *(integratie)*

Clave

1 dividendo **2** sinergia (sinergias) **3** integración

Figure 7: Automatically generated cloze exercise.

Obviously, if the user chooses an automatically generated vocabulary selection, this selection should be as accurate as possible, taking into account different parameters such as keyness and proficiency level. The further development and customization of the selection algorithms, taking into account several proficiency levels or linguistic backgrounds of the students, will boost the didactic possibilities of the tool.

Conclusion

In this paper we presented one of the first attempts to use machine learning in vocabulary selection. Although the dataset is still relatively limited, we have shown that it is a feasible task to predict students' difficulty judgments on the basis of independent variables, such as frequency in reference corpora, cognateness and prior gradation in vocabulary learning methods. The development of algorithms that define as accurately as possible the best vocabulary selection for a user of a given proficiency level is a crucial step in customizing autonomous data-driven learning initiatives, or in helping teachers to develop customized learning materials. It is our hope that these algorithms can guide both teachers and autonomous learners in their fascinating journey through the infinite world of lexis, avoiding that they feel as "amateur fishermen in the middle of the ocean" (Santos Palmou 2016: 166, our translation).

Endnotes

[1] A total number of 42 students participated in the experiment. The items were subdivided into three separate lists, so that we could dispose of 14 judgments for every item.

[2] Python difflib library.

[3] We wish to thank the authors for allowing us to digitize the index of the book publication.

[4] The feature selection method (SFFS) never obtained better results by removing any feature from the given feature subset at any given step for both machine learning algorithms.

References

Bartol Hernández, J. A. (2010). Disponibilidad léxica y selección del vocabulario. In C. Martín & V. Lagüéns Gracia (Eds.), *De moneda nunca usada: Estudios dedicados a José Mª Enguita Utrilla* (pp. 85-107). Zaragoza: Institución de Fernando el Católico.

Boulton, A. (2017). Data-Driven Learning and Language Pedagogy. In S. L. Thorne & S. May (Eds.), *Language, Education and Technology, Encyclopedia of Language and Education* (pp. 181-192). Berlin Heidelberg: Springer International Publishing.

Buyse, K., Delbecque, N., & Speelman, D. (2004). *Portavoces. Thematische woordenschat Spaans*. Mechelen: Wolters Plantyn.

Davies, M. (2005). Vocabulary range and text coverage: insights from the forthcoming Routledge frequency dictionary of Spanish. In R. Orozco, & D. Eddington, *Selected proceedings of the 7th Hispanic linguistics symposium* (pp. 106-115). Cascadilla Proceedings Project.

Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. New York: Routledge.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior research methods, 45*(4), 1246-1258.

Ferri, F. J., Pudil, P., Hatef, M., & Kittler, J. (1994). Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, *16*, 403-413.

Gabrielatos, C. & Marchi, A. (2011). Keyness. Matching metrics to definitions. In *Theoretical-methodological challenges in corpus approaches to discourse studies and some ways of addressing them*, University of Portsmouth, November 2011. Retrieved November, 20, 2018, from https://research.edgehill.ac.uk/en/publications/keyness-matching-metrics-to-definitions-2.

Goethals, P., Lefever, E., & Macken, L. (2017). SCAP-TT: Tagging and lemmatising Spanish tourism discourse, and beyond. *Ibérica*, *33*, 279-288.

Goethals, P. (2018). Customizing vocabulary learning for advanced learners of Spanish. In T. Read, B. Sedano Cuevas, & S. Montaner-Villalba (Eds.), *Technological innovation for specialized linguistic domains* (pp. 229-240). Berlin: Éditions Universitaires Européennes.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International journal of corpus linguistics*, *13*, 403-437.

Groot, P. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, *4*, 56-76.

Izquierdo Gil, M. d. C. (2005). *La selección de léxico en la enseñanza del español como lengua extranjera. Su aplicación al nivel elemental en estudiantes francófonos*. Málaga: ASELE Colección Monografías.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern recognition*, *33*(1), 25-41.

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*, 391-411.

Little, D. (2007). Language learner autonomy: Some fundamental considerations revisited. *International Journal of Innovation in Language Learning and Teaching*, *1*, 14-29.

Nation, P. (2016). *Making and Using Word Lists for Language Learning and Testing*. John Benjamins.

Navarro, J. M., & Navarro Ramil, A. J. (1996/2010). *Thematischer Grund- und Aufbauwortschatz Spanisch.* Berlin: Klett.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, *15*(11), 1119-1125.

Salazar García, V. (2004). Acercamiento crítico a la selección objetiva de contenidos léxicos en la enseñanza de E/LE. *Estudios de Lingüística, 18*, 243-273.

Santos Palmou, X. (2016). La selección del vocabulario en ELE: estado de la cuestión y nuevas metodologías. *El español como lengua extranjera en Portugal II: retos de la enseñanza de lenguas cercanas* (pp. 164-178). Ministerio de Educación, Cultura y Deporte: Subdirección General de Documentación y Publicaciones. Retrieved November 10, 2018, from http://www.educacionyfp.gob.es/portugal/dam/jcr:0698f299-82c8-4a9e-b37d-590e9fa11e8a/retos-2016-final2.pdf

Vincze, O., & Alonso Ramos, M. (2013). Incorporating frequency information in a collocation dictionary: Establishing a methodology. *Procedia-Social and Behavioral Sciences, 95*, 241-248.

Appendix

A selection of the 500+lemmas with their corresponding features and students' difficulty judgment.

POS	LEMMA	freq_abs_ref_1	freq_group_ref_1	freq_abs_ref_2	freq_group_ref_2	cognate	voc_method	keyness_ref_1	keyness_ref_2	keyness_group_ref_1	keyness_group_ref_2	Level
NC	accesibilidad	0	-1	28	4	0	3	130000	850	100	70	A
NC	accionariado	1	0	2	1	0	3	30808	16540	90	90	B-C
NC	acierto	39	4	189	5	0	1	296	-11	40	-1	B-C
NC	acorde	112	5	139	5	1	1	175	139	30	30	A-B
NC	adhesión	2	1	6	2	1	1	10717	3782	80	90	A-B
NC	adjudicatario	0	-1	4	2	0	3	130000	7388	100	90	C
NC	adversidad	18	3	5	2	0	1	930	3893	50	90	A-B
NC	ambición	112	5	77	5	1	1	286	505	30	60	A
NC	amplitud	12	3	151	5	1	1	1187	10	60	10	A-B
NC	analista	4	2	10	3	1	3	13036	5557	80	90	A
NC	ánimo	520	5	152	5	0	2	-70	9	-1	10	A
NC	anunciante	0	-1	0	-1	0	3	130000	130000	100	100	B-C
NC	aparición	297	5	170	5	0	1	-27	37	-1	10	A
NC	apoyo	320	5	175	5	0	2	885	1839	50	80	A
NC	aseguradora	3	1	5	2	1	3	6081	3893	80	90	A-B
NC	asistencia	21	3	53	4	1	2	2843	1155	70	70	A
NC	audiencia	101	5	49	4	1	1	1827	4178	70	90	A
NC	auditor	0	-1	2	1	0	3	130000	18204	100	90	A
NC	ausencia	411	5	171	5	1	2	-54	16	-1	10	A
NC	austeridad	2	1	95	5	0	1	7627	75	80	20	A
…												