ISSN 2074-9864 (online)

VIENNA ENGLISH WORKING PAPERS

HTTP://ANGLISTIK.UNIVIE.AC.AT/VIEWS/

VOLUME 22 2013

ARNE LOHMANN

Is tree hugging the way to go?
Classification trees and random forests in linguistic study

Published online: 11 June 2013

Is tree hugging the way to go?

Classification trees and random forests in linguistic study

Arne Lohmann, Vienna[*]

1. Introduction

In quantitative linguistics, recently a new type of methodological resource has become part of the empiricist's toolkit: classification trees and random forests. Particularly in the study of variation phenomena, certainly a very active area of linguistic research, trees and random forests have come to be used in addition to logistic regression modeling (Tagliamonte & Baayen 2012, Wiechmann & Kerz 2013). One of the selling points of classification trees and random forests is that they are supposedly better able to tackle certain situations where regression modeling becomes problematic due to the structure of the data. This is the case when the predictor variables are strongly correlated and/or when these variables are involved in complex interactions. In the latter case, the results of logistic regression are hard to interpret, while tree models may offer a more intuitive illustration of the results (cf. Strobl et al. 2009a: 325, Tagliamonte & Baayen 2012: 164).

This paper offers an introduction to these methods and aims at demonstrating and evaluating the new methodological resource by applying it to a dataset which is characterized by a complex interaction of the predictor variables. The case study is an analysis of the word-formation process of

[*] The author's e-mail address for correspondence: arne.lohmann@univie.ac.at

clipping, in particular the choice between back- and fore-clipping, based on data provided by Berg (2011).

The structure of the paper is as follows: In section 2 the phenomenon and the data sample are presented. Section 3 discusses a logistic regression analysis of the data conducted by Berg (2011). In section 4 classification trees and random forests are applied to the clipping sample. Section 5 discusses the results and evaluates the method.

2. The phenomenon: backclipping vs. foreclipping

The case study which I use to illustrate the application of classification trees and random forests deals with the word-formation process of clipping. More specifically, I use data by Berg (2011), who investigates the possible predictability of the choice between fore- and backclipping in English. The sample consists of 955 instances of clipping collected from dictionaries and other published sources (see Berg 2011: 4). The two clipping variants are illustrated by the following examples.[1]

(1) a. technology > tech
b. raccoon > coon

In (1a) final material of the source word is omitted, presenting an instance of backlipping, while in (1b) initial material of the source word is omitted, which thus instantiates the process of fore-clipping. The question to be empirically addressed is which factors motivate the choice between the two clipping variants. Berg (2011) identifies three possible predictor variables, viz. the lexical class, the length and the stress pattern of the source word. These variables are briefly illustrated in the following, along with the possible values they can take on:

Lexical class: whether the unclipped source word is a proper or a common noun. Values: proper noun / common noun

(2) a. Kathryn > Kath
b. caravan > van

[1] All examples in this paper are taken from Berg (2011).

Stress pattern: whether the source word bears stress on the initial syllable
Values: word-initial stress vs. non-word-initial stress

(3) a. business > biz
b. racoon > coon

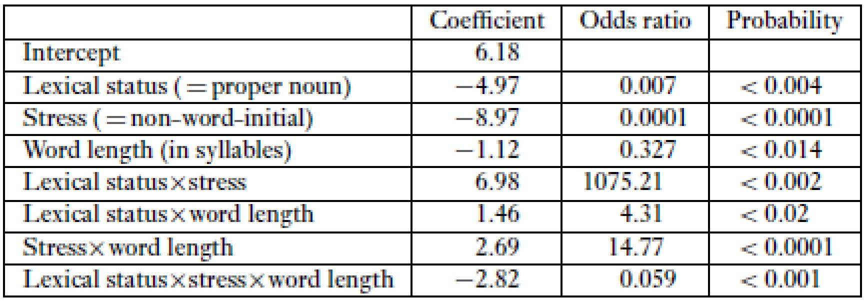
Length of source word: length of the source word in number of syllables
Values: a number ranging from 2-6, denoting the number of syllables

(4) a. Albert > Al
b. rehabilitation > rehab

Through monofactorial analysis, Berg (2011) finds that all of these variables yield an influence on the choice of clipping: Proper nouns show a higher ratio of fore-clipping as compared to common nouns. Initially stressed words show a stronger preference for back-clipping than those that are not stressed on the first syllable. And, with increasing length of the source word, the probability of back-clipping decreases.

3. A regression analysis of clipping choice

A more sophisticated analysis of the influences on clipping choice is their analysis in a multifactorial regression model, as this method takes into account the concurrent influence of the variables. Relying solely on monofactorial methods can be problematic, as interactions between the variables are not considered and it is not checked whether each of the variables yields a truly independent influence. For example, it could be that the influence of initial stress disappears, once the variable ‘lexical class’ is controlled for, as it is conceivable that proper nouns almost always bear initial stress. Berg (2011) conducts a multifactorial logistic regression analysis, the results of which are given in Table 1 below.



	Coefficient	Odds ratio	Probability
Intercept	6.18		
Lexical status (= proper noun)	−4.97	0.007	< 0.004
Stress (= non-word-initial)	−8.97	0.0001	< 0.0001
Word length (in syllables)	−1.12	0.327	< 0.014
Lexical status × stress	6.98	1075.21	< 0.002
Lexical status × word length	1.46	4.31	< 0.02
Stress × word length	2.69	14.77	< 0.0001
Lexical status × stress × word length	−2.82	0.059	< 0.001

Table 1. Results of a logistic regression analysis for back-clipping versus fore-clipping (from Berg 2011: 9)[2]

One of the prerequisites of regression is that the predictor variables must not be strongly correlated. Checking for correlations of the independent variables in the present case reveals some interdependence between them: First, in the sample initial stress wanes with increasing word length (length in number of syllables, arithmetic means of initially-stressed and non-initially stressed words: $\bar{X}_{\text{initially-stressed}}$ = 2.4, $\bar{X}_{\text{non-initially-stressed}}$ = 3.5). Second, proper nouns (which are exclusively first names in Berg's sample, see Berg 2011: 4) are on average shorter than common nouns (arithmetic means : $\bar{X}_{\text{proper nouns}}$ = 2.5, $\bar{X}_{\text{common nouns}}$ = 3.2) and third, proper nouns display a greater likelihood to be stressed on the first syllable (ratios of initial stress: proper nouns = 72%, common nouns = 55%). I therefore tested for collinearity of the predictors calculating the variance inflation factors (VIFs). All variables yield VIFs < 2, which indicates that collinearity is not a concern with this model.[3] Let us turn to an interpretation of the output of the regression analysis. The p-values in the rightmost column tell us that all variables yield significant results and are also engaged in significant interactions with each other. There is even a significant three-way interaction of all three predictors (see bottom row). Positive coefficients in the table (second column from the left) indicate a preference for back-clipping, negative coefficients a preference for fore-clipping. The coefficients are also a measure of effect size, the greater the deviation from 0, the larger the effect on clipping choice. The odds ratios (third column from the left) are an additional measure of effect size, with values between 0 and 1 indicating a preference for fore-clipping (the closer to 0, the stronger), and values above 1 indicating a preference for back-clipping

[2] Predicted odds are for back-clipping, thus positive coefficient values indicate a preference for that variant.

[3] I used the *vif* function of the *car* package in R (R Development Core Team 2011) which is able to handle categorical predictor variables (cf. Hendrickx et al. 2004: Note 4)

(the higher the value, the stronger the preference).

The main effects' negative coefficients of lexical status, stress and word length indicate that proper nouns, words with non-initial stress and longer words have a weaker preference for back-clipping. However, since all variables also feature in complex interactions with each other, an interpretation of these effects is not straightforward. With the two-way interactions an interpretation is still feasible, e.g. the positive coefficient of the interaction of lexical status with stress indicates that proper nouns are more strongly affected by stress than common nouns. However, since all variables are also involved in a three-way interaction, a plotting of the results would be required to understand the complex influences of the variables on different groups in the sample. Besides, a comparison of the effect size of the different variables as given in Table 1 is not informative, as these values are influenced by all interactions the variables take part in. In conclusion, since there is a complex interaction between the predictor variables, the data seems to lend itself well for a test of the advantages of classification trees and random forests.

4. Trees and forests of back-clipping vs. fore-clipping

This section will demonstrate the application of classification trees and random forests to the clipping data. Similar to logistic regression, these methods can be applied to predict a binary choice situation. Classification trees employ a recursive partitioning algorithm, which splits the original sample into sub-samples on the basis of the independent variables. The algorithm tests whether any of the predictor variables is significantly associated with the dependent variable (here: clipping type). If yes, the variable which exhibits the strongest association is used to create the first split, i.e. creates two sub-samples of the data which are maximally homogeneous with regards to the response variable. The resultant sub-samples are then tested again for significant associations with the predictor variables. This process continues in an iterative fashion until no further significant associations are found (see Breiman et al. 1993). The splits and resultant sub-samples are usually displayed in a tree diagram, which gives rise to the name of the method.

In a first step, I calculated a single classification tree, predicting the choice of clipping type on the basis of the variables lexical class, stress pattern and length (see above), using the *ctree* function of the *party* package in R (R Development Core Team 2011).[4] The application resulted in the following tree model (see Figure 1 below).

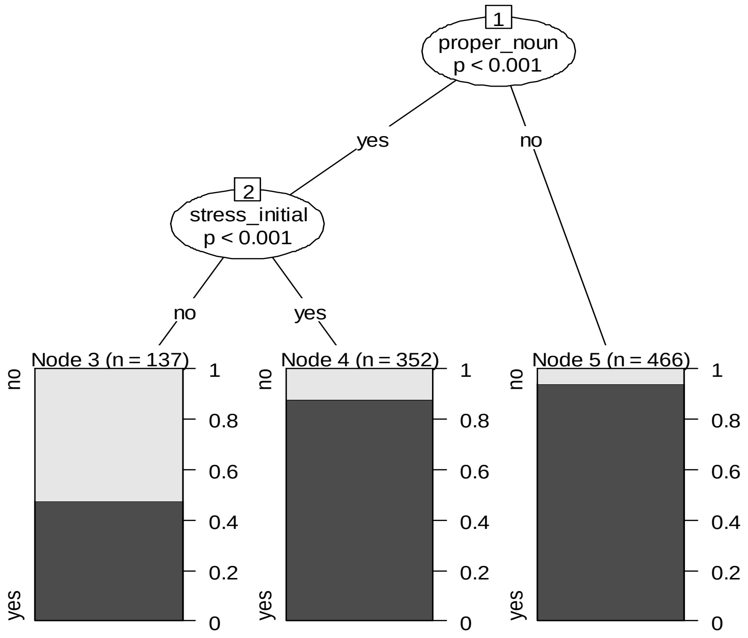


Figure 1. Classification tree of clipping choice using *ctree*

As described above, the classification tree uses the predictor variables to enforce binary splits on the data, thereby creating subsamples. The present tree features two splits, resulting in five nodes. The first split (node 1) is created by the variable ‘lexical class’ (proper noun: yes/no). It shows that with common nouns, the probability is very high that these will be backclipped (see node 5), while the probability for back-clipping is lower for proper nouns. The fact that the rightmost branch of the tree, which contains the common nouns, is not split up any further, illustrates that the influences of the other tested variables do not significantly influence clipping in that group. In contrast, if we follow the left branch starting at node 1, we see that the group of proper nouns is affected by another variable, as evident from the further split of that branch (node 2), brought about by the variable stress pattern: Initially stressed proper nouns have a significantly higher probability for back-clipping (node 4) than those which are not stressed on the initial

[4] The *ctree* function is preferred in situations in which predictor variables of different types are involved, as it ensures unbiased variable selection. This applies to the present case, as categorical variables (lexical status and stress pattern), as well as a scalar variable (length of the source word) feature in the analysis. Furthermore the *ctree* function features an internal stop-splitting mechanism, therefore no pruning is needed in contrast to other tree-growing functions, e.g. *rpart*. See Strobl et al. (2009b) on different tree-growing algorithms.

syllable (node 3). The fact that the variable 'stress pattern' leads to a split of just this one branch of the tree indicates an interaction between the variables ‘lexical status’ and ‘stress pattern’. The stress pattern of the source word matters only when dealing with proper nouns, while it has no significant influence on the clipping of common nouns. What is moreover revealed is that the variable ‘length of the source word’ does not lead to splits, thus it does not significantly improve the predictive power of the tree. With regard to the predictive accuracy of the tree we obtain an index of concordance of *C*=0.752; the tree classifies 85.7% of all data points correctly.[5]

Since the tree is based on only one particular tree-growing algorithm, it may be useful to grow a second tree for comparative purposes, using another algorithm, the *rpart* function. Unlike *ctree*, *rpart* does not feature an internal stop mechanism. It therefore calculates a maximal tree. This tree can then be cut back (‘pruned’) until only significant splits remain. This was done in the
present case using pruning methods based on cross-validation following Everitt & Hothorn (2010: 161-175). The resultant tree is illustrated below (see Figure 2).

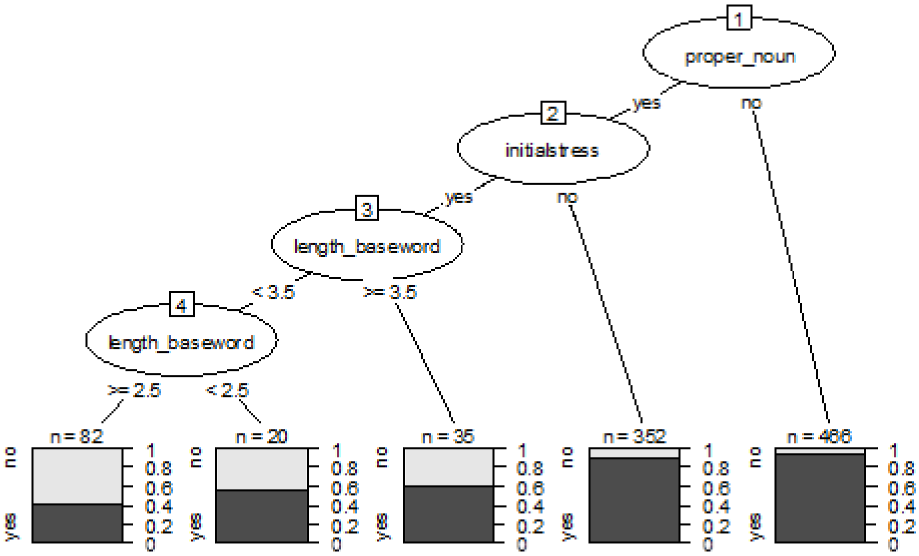


Figure 2. Classification tree of clipping choice using *rpart*

[5] The index of concordance *C* is a measure of the predictive accuracy of statistical models. It ranges from 0 to 1, with values closer to 1 denoting a high accuracy of the model.

As is immediately obvious, the tree contains more splits, as it also considers the length of the source word as a splitting variable.[6] In fact this variable produces two extra splits, as compared to the tree arrived at through *ctree*. It produces separate terminal nodes for all source words longer than three syllables (node 3), for mono- and disyllabic words, and for three-syllabic ones (node 4). The likelihood of back-clipping decreases in that order. The fact that the variable 'length of the source word' imposes a split on solely the leftmost branch of the tree, but does not affect the other two branches, is evidence for an interaction of this variable with the other two predictors. Since it affects solely those data points which are proper nouns which are initially stressed, it illustrates a three-way interaction between the predictor variables. Thus, the *rpart* tree features an even more complex interaction structure than the first tree, as the two variables stress pattern, as well as length of the source word, influence solely sub-groups in the data. Besides, in featuring more splits, the *rpart* tree yields more information than the first tree. Its predictive accuracy is slightly higher than the one of the first tree with 86.4% correct predictions (*C*=0.755).

The tree brings out an important insight on clipping as a word-formation process: While proper nouns are subject to different influences which affect the choice between back- and fore-clipping, this is not the case with common nouns which almost uniformly pick back-clipping (cf. Berg 2011).

Now that we know which variables yield a statistically significant influence on the choice between back- and fore-clipping, a further important piece of information would be the effect size of the predictor variables. Effect size is a measure of the strength of influence of an independent variable on the dependent variable. In logistic regression, common effect size measures are the coefficient values or the odds ratios of the individual variables (see Table 1 above, for a detailed discussion see Pampel 2000). However, with classification trees, effect sizes beyond the order of splits in the tree are usually not reported (cf. Tagliamonte & Baayen 2012, Wiechmann & Kerz 2013). Recall that the order of splits is determined by the strength of association between dependent and independent variables (see above). As a further measure of effect size, I suggest calculating the effect of the individual splits on the response variable. Since the independent variables are used to create binary splits which divide the original sample into sub-samples, an obvious way to measure the strength of a particular effect, manifested through

[6] Remember that *rpart* may overestimate the significance of scalar variables (cf. Note 5).

a node, would be to calculate the difference in back-clipping ratios between the two groups. The corresponding equation thus reads δ = μ_1- μ_2 (μ_1 and μ_2 denote the backclipping ratios of the two resultant sub-samples). δ may take on values between (-1) and (+1); values close to (+1) denote a strong effect towards back-clipping and values close to (-1) a strong effect towards fore-clipping. If we calculate δ for the nodes in the two trees calculated for the data, we obtain the following results, as shown in Table 2.

Tree 1 (*ctree*)	δ-value	Tree 2 (*rpart*)	δ-value
Node 1 (lexical status) *proper noun = yes*	–0.173	Node 1 (lexical status) *proper noun = yes*	–0.173
Node 2 (initial stress) *initial stress = yes*	0.403	Node 2 (initial stress) *initial stress = yes*	0.403
		Node 3 (length of source word) *>3.5 syllables*	0.168
		Node 4 (length of source word) *<2.5 syllables*	–0.148

Table 2. δ -values of nodes in the two trees for modeling clipping type

The results obtained show that it is node 2 and thereby the variable stress pattern that causes the split which differentiates most strongly between clipping types. The δ-value of 0.403 informs us that initially stressed source words have a ratio of back-clipping which is 40.3% higher than the ratio of non-initially stressed source words. Note, however, that this effect holds only for the group of proper nouns, as the relevant split is the second one of the tree, affecting only the branch containing the subsample of proper nouns. The negative value of node 1 (–0.173) yields the information that proper nouns have a back-clipping ratio which is 17.3% lower than that of common nouns. Nodes 3 and 4 illustrate the effect of source word length.

The δ-values denote how strong the effects of the different splits of the tree are and thereby indicate the different effect sizes of the corresponding variables. However, these values cannot be used for a global comparison of effect size, as some nodes only affect subsamples in the data and these effects differ from the potential effects of relevant variables on the overall sample. To illustrate that, we may compare the effect of the variable initial stress on the overall sample to its effect on the subsample of proper nouns: If initial stress was used for the first split, thus on the entire sample, the δ-value would be solely 0.165, as compared to 0.403 when applied to solely proper nouns, as in the trees calculated. In evaluating effect size, it is therefore a good idea to take into account two pieces of information: the order of splits as an indicator of global effect size and the δ-values of individual nodes indicating strength of effect on subsamples in the data.

The calculation of the two trees naturally raises the question of whether the more detailed tree based on the *rpart* algorithm is really justified, or whether the more parsimonious *ctree* tree is more accurate. One way to arrive at an answer to this question is to calculate an ensemble of trees, through the application of the technique termed 'random forests'. Random forests have the advantage that they yield more stable results than individual trees, as they are not as sensitive to particular characteristics of the individual sample. With single classification trees it can happen that only small distributional changes in the sample employed result in very different tree structures, an issue which is also known as the instability of single trees (Strobl et al. 2009a). This can be a problem since the tree may yield a result for the relation between variables which is true of the sample, but not of the population. The reason for why classification trees react sensitively to the random variability of the sample lies in the hierarchical splitting mechanism underlying their growth, as every split influences the further growth of the tree as a whole. If, for instance, the first split is chosen because of a non-representative characteristic of the sample, it deteriorates the quality of the whole tree (see Strobl et al. 2009a: 330). Since the essential task of inferential statistics is to license statements about the population, this is a potentially serious issue.

In order to mitigate that shortcoming, forest methods grow many trees over different samples which are created from the original sample and then calculate an average over the ensemble of trees. The technique aims at a certain diversity of the trees which has been shown to improve the predictive power of the forest as a whole (see Strobl et al. 2009a). This diversity is implemented through two features: resampling and random selection of variabes. First, for every tree of the forest two subsamples are created. One random subsample which is used to grow the classification tree (= the learning sample), whose validity is then tested against those datapoints which were not included (= the test sample). Second, during the growth of the individual trees for every possible split only a random subset of all available predictor variables is used. In the present case, where we test the influence of three predictor variables, for every possible split only two out of three variables are tested. Through this random selection of variables even more diverse trees are grown (see Strobl et al. 2009a: 333). For the case study of clipping, a random forest consisting of 3,000 trees was grown, testing the three predictor variables mentioned above.[7]

Since it is usually not feasible to manually inspect and evaluate 3,000 individual trees, a further calclulation is necessary to assess the importance of the individual variables. One way of doing this is through ‘conditional permutation’ of the predictor variables (see Strobl et al. 2008, Tagliamonte & Baayen 2012: 160-162). This means that the values of potentially important variables are randomly altered (permuted). This permutation creates a vector of values which has no association with the dependent variable anymore. It is then tested how severe the loss in predictive accuracy of the forest is when this permuted version of the variable is employed instead of the original one. If the permutation results in a considerable loss of predictive power, this is indication that we are dealing with an empirically relevant variable which has a high importance for the classification of the response variable. Conversely, if the model hardly deteriorates through inclusion of the permuted variable, the original variable is most likely not very important. Let me exemplify this process: With the variable ‘lexical class’, we know that the value ‘common noun’ is associated with a high ratio of back-clipping in contrast to the value ‘proper noun’. Through permutation, the values proper noun/common noun are randomly mixed, and a new permuted version of the variable ‘lexical class’ is created. Through the random permutation there should no longer be an association between back-clipping and the value ‘common noun’. This permuted version of the variable ‘lexical class’ is now used to grow trees along with the other un-permuted variables and the deterioration in accuracy of prediction is calculated. The results of this calculation are presented in the barplot below.[8] The lengths of the bars indicate the importance of the corresponding variables for predicting the choice between back- and fore-clipping. The dotted line indicates statistical significance, those variables whose bars surpass it contribute statistically significantly to the overall accuracy of the forest.[9]

[7] The R function *cforest* of the party package was used as it is better suited for samples with variables of different types than the alternative algorithm randomForest (see Strobl et al. 2009b: 17).

[8] The *varimp* function of the *partykit* package was used (see Strobl et al. 2008). The employed parameter settings were ntree=3000 and mtry = 2, seed set at 147. Several forests were calculated, with varying the parameter settings. The rankings of the variables proved to be stable across several runs of the algorithm. Results are displayed in Figure 3.

[9] One way to calculate whether a variable contributes significantly to the forest is to compare it to those variables whose influence has been shown to be detrimental to overall model performance, which would shows as a negative vector in the barplot. The lowest negative value is taken as an absolute value to which the positive contributions are compared. Only those variables which surpass this value are

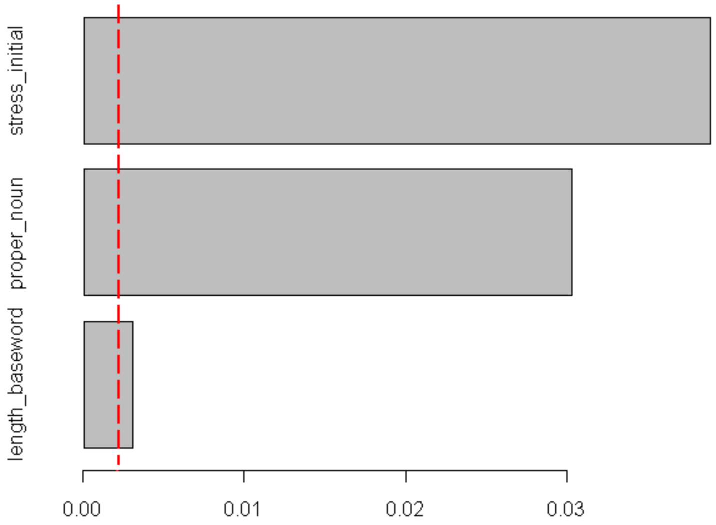


Figure 3. Variable importance (random forest for clipping data)

The results show that the three tested variables significantly influence the choice between the two clipping types. Comparing their importance shows that lexical status, as well as the stress pattern of the source word yield important influences on the prediction of clipping type, with the latter being the most important variable. The length of the source word is the variable which is least important.

In informing us about the statistical significance and relative importance of the three variables the calculation also answers the question which of the two trees reported above does better justice to the structure in the data: It is the more elaborate *rpart* tree, as it uses the variable 'length of the source word' to create splits, whose significance is confirmed by the application of random forests. Since the variable is however of only lesser importance, also the *ctree* tree model is acceptable.[10]

Overall, the results for variable importance mirror the effect sizes of the individual nodes brought about by the respective variables: The stress pattern of the source word is the variable which brings about the node with the largest δ-value and it is also the one which has the greatest importance for clipping

considered to contribute significantly to the forest (see Shih 2011, Strobl et al. 2009a: 342). In the present case, however, there was no variable which yielded a negative importance score. Therefore I introduced a binary random variable in order to simulate a variable which has no relation with the dependent one in order to produce such a negative importance score.

[10] The minor importance of the variable ‘length of the source word’ is also reflected in a similar predictive accuracy of the two trees

classification as a whole (see Table 2 and Figure 3). Note however, that the two values do not measure the same thing. While the δ-value denotes the effect of a certain split, thus the effect size of a variable on a certain subsample in the data, variable importance denotes the overall contribution to predictive accuracy. A variable may thus have a large δ-value in strongly affecting the distribution of certain datapoints, but it may be an overall unimportant one, if it is active only in a small subset of the data. Often however, as is also the case here, variables score high on both dimensions.

Finally, the random forest can also be used to predict the values of the independent variable. This is arrived at through a 'voting' system, where for each datapoint the outputs of all individual trees are calculated. The value which receives the most votes is the output of the forest as a whole. This 'voting' technique results in a predictive accuracy of *C*=0.81 and 87.8% correctly classified data points. These values represent an improvement over the individual trees (see above), hence, the random forest makes slightly better predictions than the individual trees.

5. Discussion and conclusion

Classification trees and random forests were successfully applied to the present case study of clipping choice. In the following I will discuss and evaluate the application of the methods.

One assumption was that the new tools can better handle complex interactions between the predictor variables. Above, we saw that with the clipping data, logistic regression yielded results which were hard to interpret (see Table 1). In contrast, the tree models calculated provide a straightforward illustration of the interaction effects. Their architecture allows for an easy interpretation of possible interactions, as we simply have to compare the different branches of the tree: If a certain variable is used to create splits in one branch of the tree, but not another, it influences only a subsample, thus interacts with the variable which created the branch. In the present case study, the trees brought to light an interesting result on the word-formation process of clipping: While the choice between clipping types is subject to different influences in the case of proper nouns, these variables do not significantly influence common nouns, which are almost uniformly backclipped. It needs to be mentioned, however, that with very elaborate trees which are based on a large number of predictor variables, the interpretation of a classification tree could also become more complex than in the present case (cf. Strobl et al. 2009a: 328-329).

I furthermore calculated a ‘random forest’ which is an ensemble of classification trees. Since single trees are not always reliable, as they are sensitive to characteristics of the sample, random forests calculate which variables have a significant impact on the dependent variable with varying samples. Moreover, random forests provide the user with further measures of variable importance (see Figure 3). Lastly, also the forest can be understood as a model which predicts the values of the dependent variable. Thus, in the present case, instead of relying on one tree, we can also ‘ask’ the forest to predict the choice between the two clipping variants and, as we have seen, the predictive accuracy of the random forest is higher than that of the individual trees.

As the random forests also allow for predicting the values of the dependent variable, it seems that it can be treated like other statistical models, for example logistic regression. However, there are important differences between these two methods: First, every tree in the forest is based on a random choice of variables, which is calculated on the basis of a random sample (see above). Therefore the results of a random forest application vary with every single calculation (see Strobl et al. 2009, Shih 2011). Thus, at least theoretically, the results for importance and statistical significance of the tested variables are not stable across different runs of the random forests function. This characteristic obviously has a negative effect on the test-retest reliability of the method. In actual practice, however, this characteristic will not constitute a major problem in most cases, due to the fact that when a large ensemble of trees is grown, there should be a regression towards the mean of the variable values. Thus the actual differences between two different calculations will only be slight, given the forests are large enough. However, this characteristic still distinguishes this method from regression analyses: In a regression equation the coefficients and measures of variable importance are invariable and will be the same across any two calculations, as long as we apply the same regression function and follow the same procedures during model fitting.[11] With random forests the user is advised to keep the potential computational instability in mind, and to always calculate several random

[11] In actual practice, the distinction between the two techniques may not bet that sharp, as during model fitting very often more than one model is calculated, for instance if bootstrapping is applied, during which often more than 100 different models are calculated based on different samples (see e.g. Baayen 2008). This would then be quite comparable to a random forest in also being a sort of *ensemble* technique. It would therefore be more adequate to compare a single regression model to one particular tree and a random forest to separate regression models over many different sub-samples. Thus the difference between the two methods lies more in their application than in their mathematical characteristics. Nevertheless, the aim of a regression model is always to arrive at one regression equation, which is not the case with random forests.

forests to verify that the results are stable across a number of runs (Shih 2011: 4).

References

Baayen, Rolf H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Univ. Press.

Berg, Thomas. 2011. “The clipping of proper and common nouns.” *Word Structure* 4(1). 1-19.

Breiman, Leo; Friedman, Jerome; Stone, Charles J.; Olshen, R.A. 1993. *Classification and regression trees* (The Wadsworth statistics, probability series). New York: Chapman & Hall.

Everitt, B. S.; Hothorn, Torsten. 2010. *A Handbook of Statistical Analysis Using R*. Boca Raton, FL: Chapman & Hall/CRC.

Hendrickx, J; Belzer, B.; Grotenhuis M.; Lammers, J. 2004. “Collinearity involving ordered and unordered categorical variables.” paper given at the RC33 conference in Amsterdam August 2004. http://www.belgeler.com/blg/2a5r/collinearity-with-categorical-variables (April 29, 2013).

Pampel, Fred. 2000. *Logistic Regression: A Primer* (Quantitative Applications in the Social Sciences 132). Thousand Oaks, CA: Sage University Papers.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shih, Stephanie. 2011. “Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide.” www.stanford.edu/~stephsus/R-randomforest-guide.pdf. (March 28 2011)

Strobl, Carolin; Boulesteix, Anne-Laure; Kneib, Thomas; Augustin, Thomas; Zeileis, Achim. 2008. “Conditional variable importance for random forests.” *BMC Bioinformatics* 9(1). 307.

Strobl, Carolin; Malley, James; Tutz, Gerhard. 2009a. “An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests.” *Psychological Methods* 14(4). 323-348.

Strobl, Carolin; Hothorn, Torsten; Zeileis, Achim. 2009b. “Party on!” *The R Journal* 1(2). 14-17.

Tagliamonte, Sali A.; Baayen, Rolf H. 2012. “Models, forests and trees of York English: Was/were variation as a case study for statistical practice.” *Language Variation and Change* 34, 135-178.

Wiechmann, Daniel; Kerz, Elmar. 2013. “The positioning of concessive adverbial clauses in English. Assessing the importance of discourse-pragmatic and processing-based constraints.” *English Language and Linguistics*, 17(1), 1-23.

How to contact VIEWS:

VIEWS c/o
Department of English, University of Vienna
Spitalgasse 2-4, Hof 8.3
1090 Wien
AUSTRIA

fax	**+ 43 1 4277 9424**
e-mail	**views.anglistik@univie.ac.at**
W^3	**http://anglistik.univie.ac.at/views/** **(all issues available online)**

Impressum:
Eigentümer, Herausgeber & Verleger: VIEWS**,** c/o Institut für Anglistik & Amerikanistik der Universität Wien, Universitätscampus AAKH, Spitalgasse 2, 1090 Wien, Austria**. Für den Inhalt verantwortlich**: Andreas Baumann, Arne Lohmann. **Redaktion:** Christiane Dalton-Puffer, Nora Dorn, Malgorzata Fabiszak, Helen Heaney, Gunther Kaltenböck, Kamil Kaźmierski, Evelien Keizer, Angelika Rieder-Bühnemann, Nikolaus Ritt, Anita Santner-Wolfartsberger, Barbara Seidlhofer, Claudio Schekulin, Barbara Soukup, Ute Smit, Henry G. Widdowson, Eva Zehentner. **Alle:** c/o Institut für Anglistik & Amerikanistik der Universität Wien, Universitätscampus AAKH, Spitalgasse 2, A - 1090 Wien. **Herstellung:** VIEWS