A forest of possibilities: Predicting the choice of *can*, *could*, *may*, *might* and *be able to*

Ilse Depraetere[1], Bert Cappelle[1], Ludovic De Cuypere[2], Cyril Grandin and Benoît Leclercq[3],
[1]University of Lille, [2]Ghent University, [3]University of Paris 8, {ilse.depraetere|bert.cappelle}@univ-lille.fr, Ludovic.DeCuypere@UGent.be, cyril.grandin@live.fr, benoit.leclercq04@univ-paris8.fr

On what basis do speakers of English choose between the modal verbs that can express possibility? Various factors determining their choice have been mentioned in the literature (see, *inter alia*, Coates 1995, Collins 2009, Depraetere and Langford 2020, Huddleston and Pullum et al. 2002, Leclercq and Depraetere 2022, Palmer 1987), but what makes it difficult to predict the use of one modal over another is that these factors can interact in intricate ways. Therefore, rather than looking at the impact of each potential factor on its own, we need to consider all these factors together, with an eye to discovering the most impactful ones and the major patterns of interactions among them.
We extracted 500 occurrences of each of the five possibility modals from COCA (Davies 2008-2019) and annotated the set of 2500 utterances for 31 predictor variables, including the semantic category of the modal (epistemic possibility, ability, permission, etc.), its temporal relation with respect to the complement situation (simultaneity, anteriority, posteriority), voice of the clause (active, passive), animacy of the subject, polarity (negative, contracted *not*, positive), speech act (assertive, non-assertative), and the genre of the text in which the modal appeared (academic, fiction, etc.).
To examine the multivariate effect of these semantic, morpho-syntactic, pragmatic and discursive factors, we fitted Conditional Inference Trees (CITs) using the ctree() function (Hothorn, Hornik and Zeileis 2006) in R (R Core Team 2021). In addition, we built Conditional Random Forest models (CRFs), using cforest()(Hothorn et al. 2006, Strobl et al. 2007, 2008). This allowed us to establish the variable importance in multiple subsamples of the data. We fitted a general model for all five modals as the outcome variable, as well as three more specific ones, zooming in on (i) *be able to* vs. *can* vs. *could*, (ii) *may* vs. *might*, and (iii) the former three vs. the latter two.

Our findings, reported on in Depraetere, Cappelle, Hilpert et al. (2023: Ch. 3), show that the semantic category of the modal is most discriminating, with epistemic modality clearly favouring *may* and *might* and non-epistemic ('root') modality predicting the choice of the three other modal expressions. The temporal location of the modality is the next most significant factor. *May* and *might* (Fig. 1) can themselves be discriminated by genre (academic discourse favouring the former) and then by temporal location: modality situated in the past almost exclusively favours *might*, while for modality in the present, we again need to consider the semantic category to predict *may* vs. *might* (the two being fairly equally likely, except for what we call 'general situation possibility' and opportunity, which favour *might*). The other three modals can likewise be predicted on the basis of semantic category, temporal location and actualization (Fig. 2).

Our analysis has implications for language description, in that it shows which factors do and do not play a significant role and how the significant ones present nested conditions ('if (if…)'). This in turn can be useful for pedagogy, in that it suggests 'ideal' example sentences that simultaneously instantiate the relevant conditions – constructions, in a sense – for the possibility modals.

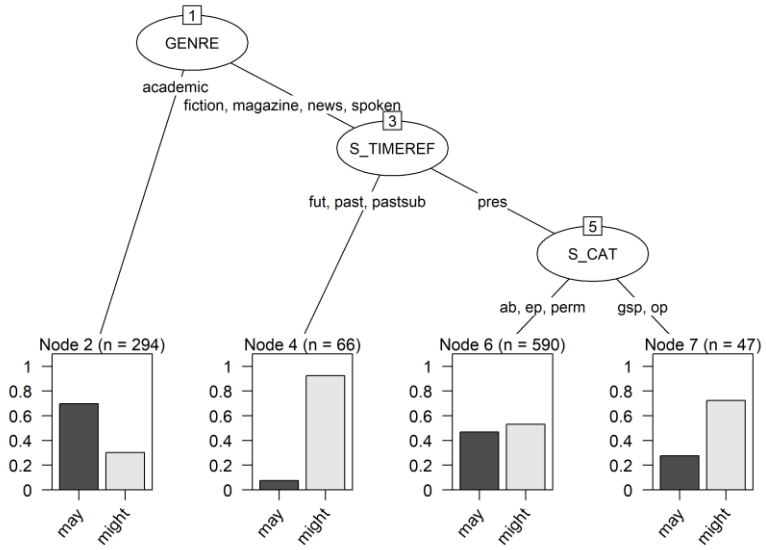

Fig. 1: CIT for the binary outcome *may* vs. *might* (S_TIMEREF = the semantic variable temporal location; pastsub = past tense in subclause triggered by a past tense in the main clause; S_CAT = semantic category; ab = ability, ep = epistemic, perm = permission, gsp = general situation possibility, op = opportunity; n = number of occurrences in a 'bin')

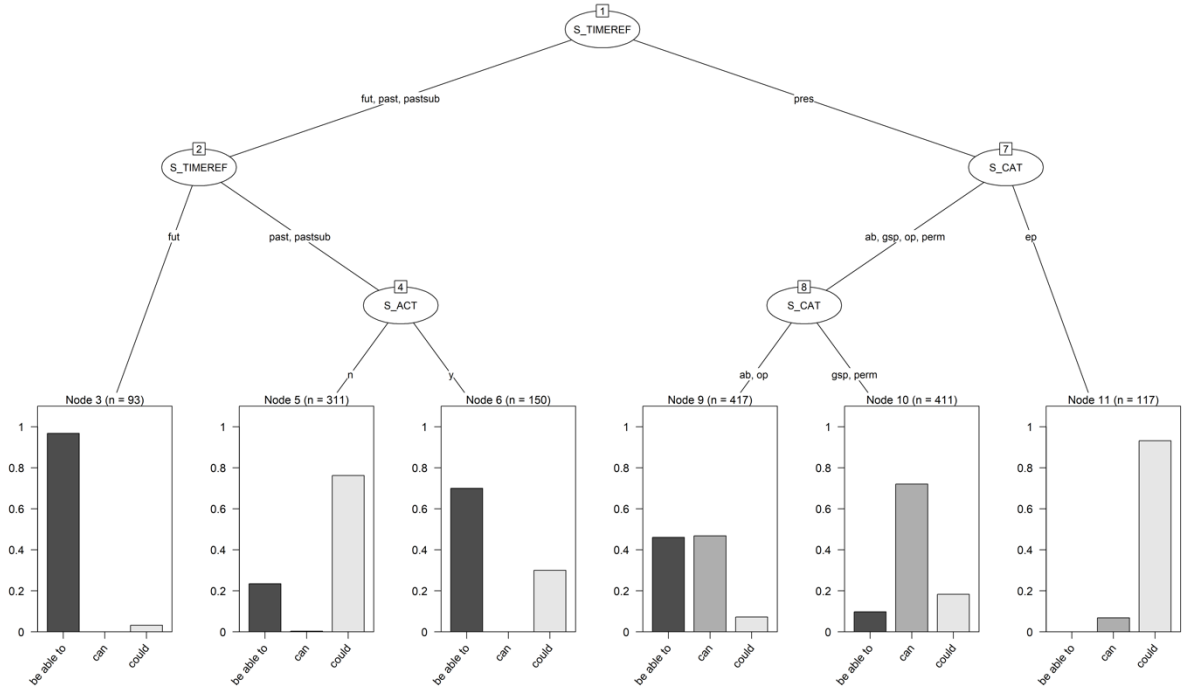

Fig. 2: CIT for the triadic outcome *be able to* vs. *can* vs. *could* (S_TIMEREF = the semantic variable temporal location; pastsub = past tense in subclause triggered by a past tense in the main clause; S_ACT = the semantical variable actualization; y/n = yes/no; S_CAT = semantic category; ab = ability, ep = epistemic, perm = permission, gsp = general situation possibility, op = opportunity; n = number of occurrences in a 'bin')

References

Coates, Jennifer. 1995, Root and epistemic possibility in English. In: Bas Aarts and Charles F. Meyer (eds.), *The Verb in Contemporary English: Theory and Description*, pp. 145-156. Cambridge: Cambridge University Press.

Collins, Peter. 2009. *Modals and Quasi-modals in English*. Amsterdam and New York: Rodopi.

Davies, M. (2008–2019). *The Corpus of Contemporary American English: 600 million words, 1990-present*.

Depraetere, Ilse, Bert Cappelle, Martin Hilpert et al. 2023. *Models of Modals: From Pragmatics and Corpus Linguistics to Machine Learning*. Berlin and New York: De Gruyter Mouton.

Depraetere, Ilse and Chad Langford. 2020. *Advanced English Grammar: A Linguistic Approach*. Second edition. London: Bloomsbury Academic.

Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro and Mark Van Der Laan. 2006. Survival Ensembles. *Biostatistics*, 7(3), 355-373.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Huddleston, Rodney, Geoffrey K. Pullum, et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Leclercq, Benoît and Ilse Depraetere. 2022. Making meaning with *be able to*: modality and actualization. *English Language and Linguistics*, 26(1), 27-48.

Palmer, Frank. 1987. *The English Verb*. Second edition. London: Longman.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).