Rethinking English modal constructions: From feature-based paradigms to usage-based probabilistic representations

Book of Abstracts

Workshop organizers:
Ilse Depraetere
Bert Cappelle
Martin Hilpert

A new model for the analysis of modal meaning: a theoretical bridge between Construction Grammar and Relevance theory

Benoît Leclercq

Even a quick investigation of the literature reveals the difficulty to apprehend the use and function of modals, which often serve as test ground for various empirical and theoretical hypotheses (cf. Nuyts & van der Auwera 2016). In spite of extensive research, there is as yet no consensus on what exactly defines the semantic category of modality and whether modals are monosemous or polysemous, and it remains a challenge to identify the main co(n)textual factors that contribute to shaping the meaning of modal sentences. Answers to these questions will vary depending on the theoretical background in which they are couched. The aim of this paper is to look at the benefits of combining insights from Construction Grammar (Hoffmann & Trousdale 2013, Hilpert 2019) and Relevance Theory (Sperber & Wilson 1995, Clark 2013) to deal with these issues and to present a new model for the analysis of modal meaning.

Previous constructionist and relevance theoretic approaches to modality will be critically discussed (Groefsema 1995, Papafragou 2000, Wärnsby 2002, 2016, Boogaart 2009), with an eye to identifying their respective strengths and weaknesses. The new model will then be put forward. First, the category of modality will be defined and it will be argued that the complexity of pinning down the meaning of modal expressions comes from the dual nature of their semantic content, both conceptual (i.e. possibility/necessity) and procedural (i.e. the speaker's attitude, root or epistemic), a distinction central to Relevance Theory (cf. Escandell-Vidal, Leonetti & Ahern 2011). This observation will be shown to be consonant with Hilpert's (2016, 2019) claim that modals are semi-schematic constructions (i.e. *can* V-INF, *must* V-INF, *may* V-INF, etc.), which from a constructionist viewpoint entails that they are semantically hybrid (Traugott & Trousdale 2013: 13). Second, it will be argued that while modal constructions are semantically polysemous (as assumed in Construction Grammar), the interpretation of modal sentences remains essentially a pragmatic affair guided by general cognitive principles (as assumed in Relevance Theory). This semantically-guided pragmatic process will be discussed in terms of *lexically-regulated saturation* (Depraetere 2010, 2014, Leclercq 2019). This paper will then end with the observation that the interpretation of modal sentences can sometimes be directly affected by the use of more fixed modal idioms (e.g. *I*

can't help but VP) whose functional properties contribute to the utterance differently from that of the modal constructions found inside them (e.g. *can* V-INF).

Boogaart, R. (2009). Semantics and pragmatics in Construction Grammar: The case of modal verbs. In A. Bergs and G. Diewald (Eds.), *Contexts and constructions*, 213-41. Amsterdam/Philadelphia: John Benjamins.

Clark, B. (2013). *Relevance Theory*. Cambridge: Cambridge University Press.

Depraetere, I. (2010). Some observations on the meaning of modals. In B. Cappelle & N. Wada (Eds.), *Distinctions in English grammar, offered to Renaat Declerck*, 72-91. Tokyo: Kaitakusha.

Depraetere, I. (2014). Modals and lexically-regulated saturation. In *Journal of Pragmatics* 7: 160-77.

Escandell-Vidal, V., M. Leonetti & A. Ahern. (2011). *Procedural Meaning: Problems and Perspectives*. Bingley: Emerald Group Publishing.

Groefsema, M. (1995). *Can*, *may*, *must* and *should*: a relevance theoretic account. In *Journal of Linguistics* 31: 53-79.

Hilpert, M. (2016). Change in modal meanings: Another look at the shifting collocates of *may*. In *Constructions and Frames* 8 (1): 66-85.

Hilpert, M. (2019). *Construction grammar and its application to English*. (2nd Ed.) Edinburgh: Edinburgh University Press.

Hoffmann, T. & G. Trousdale. (2013). *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press.

Leclercq, B. (2019). *On the semantics-pragmatics interface: a theoretical bridge between Construction Grammar and Relevance Theory*. PhD thesis. Université de Lille.

Nuyts, J. & J. van der Auwera. (2016). *The Oxford handbook of modality and mood*. Oxford: Oxford University Press.

Papafragou, A. (2000). *Modality: Issues in the semantics-pragmatics interface*. Amsterdam: Elsevier.

Sperber, D. & D. Wilson. (1995). *Relevance: Communication and cognition*. (2nd Ed.) Oxford: Blackwell.

Traugott, E. C. & G. Trousdale. (2013). *Constructionalization and constructional changes*. Oxford: Oxford University Press.

Wärnsby, A. (2002). Modal constructions? In *The Department of English in Lund: Working Papers in Linguistics,* 2.

Wärnsby, A. (2016). On the adequacy of a constructionist approach to modality. In *Constructions and Frames* 8: 40-53.

Possibility modals: which semantic and pragmatic conditions make them possible?

Cyril Grandin, Bert Cappelle and Ilse Depraetere

This presentation reports on the findings from an extensive corpus-based study that aims to explore pin down the subtle (syntactic, semantic and pragmatic) differences between five possibility modals: *can*, *could*, *may*, *might* and *be able to*. A sample of 2500 tokens from the

Corpus of Contemporary American English (Davies 2008-) were manually annotated in terms of 36 variables that previous smaller-scale theoretical and empirical studies have argued to be relevant to modal verb selection. Some of these variables are mainly semantic in nature, for example, whether the subject of the modal expression is generic or not (e.g., *Potatoes can be planted anytime in summer* vs. *Once the leaves turn pale yellow and wilt I can harvest potatoes*). Others are syntactic. Logistic regression techniques will be used to determine which of these are significant variables for the speaker's selection of a particular modal expression.

Davies, Mark. (2008-). The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at http://www.english-corpora.org.

Necessity modals and the role of source as a predictive factor

Ilse Depraetere, Bert Cappelle, Cyril Grandin, Benoît Leclercq

The distinction between objective vs. subjective modality is one that is standardly addressed in empirical studies on modal necessity verbs that aim to predict the choice of modal: *must* is commonly argued to be more subjective than *have to*, for instance (Huddleston & Pullum *et al*. 2002: 205–6, Quirk *et al*. 1985: 225, Verstraete 2001).

A major issue is obviously that of making a semantico-pragmatic feature of this type operational for corpus annotation. In Depraetere and Verhulst (2008) the subjective vs. objective distinction is made explicit in terms of modal source and a four-class taxonomy (subjective/discourse-internal, circumstantial, conditional, rules and regulations) is put forward, *must* and *have to* serving as a test case.

A further challenge is that of considering a wide range of modal verbs, rather than zooming in on the choice of one in a set of two verbs.

The aim of this presentation is to lay bare significant patterns in the distribution of sources in sentences with the five main necessity verbs (*must*, *have to*, *should*, *ought to* and *need*). A sample with 2500 examples randomly selected from COCA has been annotated in terms of five sources : discourse-internal (speaker/hearer), circumstantial, rules and regulations, conditional and subject-internal sources. Statistically significant correlations between specific verbs and specific sources will be presented and discussed, also in view of hypotheses put forward in the previous work.

Depraetere, Ilse and An Verhulst. 2008. Source of the modality: a reassessment, *English Language and Linguistics* 12: 1–25.

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

Quirk, Randolph, Greenbaum Sidney, Leech Geoffrey and Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman .

Verstraete, Jean-Christophe, 2001. Subjectivity and objectivity: interpersonal and ideational functions in the English modal auxiliary system. *Journal of Pragmatics* 33, 1505-1528.

Modals in the network model of Construction Grammar

Suzanne Flach and Martin Hilpert

In this presentation, the authors move away from the feature-based paradigms implemented in the first two presentations. It is hypothesized that knowledge of modal expressions is exemplar-based and probabilistic. In other words, speakers' knowledge of modal expressions is not to be modeled as a paradigm of forms that can be fully described through a set of cross-cutting categorical features, but rather as a network of form-meaning pairs (Hilpert 2014, Hilpert and Diessel 2016) in which the forms of modal expressions are connected to a range of meanings through associative links. Differences in association strength account for the fact that speakers choose a certain modal expression in a certain speech situation. Speakers' knowledge of modal expression is not viewed as a discrete one-to-one mapping between a form and a list of semantic features, but rather as knowledge of the probability that a given form will convey a certain meaning in a certain context. The central methodological instrument will be semantic vector space modeling (Turney and Pantel 2010). The method is based on the hypothesis that linguistic elements that occur in similar contexts will be semantically related. Hilpert (2012) creates a semantic vector space of the nine core English modal auxiliaries and shows on the basis of diachronic data that the modal *may* has undergone a semantic shift towards more epistemic meanings. Hilpert (2016) follows up on this result and identifies the contextual elements that are chiefly responsible for the diachronic semantic shift of *may*. Changing co-occurrence frequencies in the COHA reveal that *may* has come to be used more often with verbs that are abstract, stative, and unrelated to animate subjects, such as *depend*, *exist*, *involve*, or *indicate*. This presentation builds on these results and conduct parallel analyses for two pairs of expressions: *should* vs. *ought to*, and *can* vs. *be able to*. For each pair, semantic vector spaces will be constructed in order to investigate the distributional differences between expressions that partly overlap in their respective meanings. Whereas in Hilpert (2016), only verbal complements were used to define the dimensions of the vector space, the vector spaces in the present project will integrate other contextual features, such as the subject, polarity, and adverbial modification. As an additional refinement, we will compare different types of dimensionality reduction techniques: we will compare Multidimensional Scaling (MDS) that is used in Hilpert (2016), to standard Principal Component Analysis (PCA) or its sparse variant Sparse PCA (Zou et al. 2006).

You must/have to choose – how speakers decide between near-synonymous modals. Convergent evidence from psycholinguistics

Suzanne Flach, Bert Cappelle and Martin Hilpert

This presentation investigates the factors that underlie speakers' behavior when they choose between near-synonymous modal auxiliaries. When and why does a speaker say *You have to choose* instead of *You must choose*? We investigate this question through an experimental study in which respondents indicate their relative preference for one of two stimuli.

Our study is based on a set of corpus-based studies of alternative modal auxiliaries, which examine the structural and semantic features that distinguish between modals such as *may* and *might*. Flach (2020) analyzes the impact of collocating adverbs; Grandin (to appear) uses regression modeling in order to test for the influence of factors such as aspect, voice, and syntactic constituency; Hilpert and Flach (to appear) apply distributional methods in order to determine the lexical contexts that reliably distinguish between alternative modals. The results of the three studies indicate that the choice between *may* and *might* is driven, amongst other factors, by lexical collocates, the distinction of singular vs. plural subjects, subject animacy, the clause type that contains the modal, as well as the semantic domain of the surrounding vocabulary.

Corpus-linguistic methods are usefully complemented by psycholinguistic techniques because the former often generate hypotheses that can only be tested under controlled conditions (Bresnan 2007, Bresnan & Ford 2010). The present paper tests the results of the corpus-based studies through the paradigm of 100-split tasks, in which speakers have to assign relative preference scores to alternative, near-synonymous modal expressions. Stimuli are constructed so that they incorporate significant variables of the corpus-based studies in a cross-cutting way. To illustrate, the examples below offer choices between *must* and *have to* that differ with respect to subject number and syntactic context.

	must	*have to*
Singular subject, If-clause	*I wonder if she must hand in the paper by tomorrow.*	*I wonder if she has to hand in the paper by tomorrow.*
Plural subject, If-clause	*I wonder if they must hand in the paper by tomorrow.*	*I wonder if they have to hand in the paper by tomorrow.*
Singular subject, That-clause	*That she must hand in the paper by tomorrow is clear.*	*That she has to hand in the paper by tomorrow is clear.*
Plural subject, That-clause	*That they must hand in the paper by tomorrow is clear.*	*That they have to hand in the paper by tomorrow is clear.*

Participants compare stimuli of the same row. Two perfectly equivalent stimuli receive 50 points each; in cases of preference asymmetries, the preferred stimulus receives up to 100 points. The resulting scores indicate whether factors such as subject number or syntactic context influence participants' preferences and thereby allow us to test the psychological reality of the corpus-based studies.

On the basis of our results, we argue that 100-split tasks that incorporate corpus-derived findings can not only provide insights into modality or any other grammatical phenomenon, but that they furthermore allow researchers to synthesize, compare, and evaluate claims that previous studies have advanced.

Bresnan, J. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (eds.), *Roots: Linguistics in Search of Its Evidential Base*. Berlin: Mouton de Gruyter, 77-96.

Bresnan, J. and M. Ford. 2010. Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English. *Language* 86(1): 186-213.

Flach, S. (2020). Beyond modal idioms and modal harmony: A corpus-based analysis of gradient idiomaticity in MOD + ADV collocations. *English Language and Linguistics*, 1-23. doi:10.1017/S1360674320000301

Grandin, Cyril. To appear. *A multifactorial analysis of the meaning of modal verbs*. PhD. Lille University.

Hilpert, Martin and Susanne Flach. 2020. Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities*, fqaa014, https://doi.org/10.1093/llc/fqaa014

Revisiting modal sense classification with state-of-the-art language models

Mathieu Dehouck and Pascal Denis

Understanding and modeling the use of modal verbs is an important topic in both linguistics and natural language processing with applications in sentiment analysis or fact checking. In this paper, we are interested in disambiguating the use of English modal verbs as an automatic prediction task. That is, given a modal verb in context, we want to predict the intended sense of the verb.

This problem was first addressed by Ruppenhofer and Rehbein (2012) who produced the first annotated corpus for Modal Sense Classification (MSC). They then went on predicting modal sense on their freshly annotated corpus using hand crafted syntactic features. This first data-set suffered strong data imbalance with sense distribution being very skewed. Zhou et al. (2015) thus proposed to annotate a new corpus via cross-lingual modal sense projection from German sentences. They also introduced a new set of semantic features, and compared the performance of their new features with those from Ruppenhofer and Rehbein on their new balanced data-set.

In 2019, Li et al. proposed to complement previously hand-crafted features with automatically learned word embeddings (vectorial representations of words learned from large amounts of text). Their results showed improvement for some modal verbs (*can* and *may*) compared to previous method. However, their mitigated results could be explained by the limited amount of training data available.

In this paper, we revisit MSC using more recent word representation methods, namely neural transformers (BERT). These new representations are learned on the basis of huge amounts of text with an objective function that infuses them with their context, and are then tailored at run time to best represent each token. We show that models based on these new contextual representations achieve results on a par with those based on previously hand-crafted features while using only information extracted automatically from unannotated data.

We then look at self-training techniques in an attempt to alleviate the small amount of training data. We explore both self-training proper and ensemble-voting as a way to select automatically annotated data to be added to the training set.

We also apply our models to a newly annotated corpus of English modal verbs that uses a different annotation scheme than previous corpora and report results for this new corpus.

Bo Li, Mathieu Dehouck and Pascal Denis. *Modal sense classification with task-specific context embeddings*. ESANN 2019, Bruges, Belgium.

Mengfei Zhou, Anette Frank, Annemarie Friedrich, and Alexis Palmer. Semantically Enriched Models for Modal Sense Classification. In *Proceedings of the EMNLP Workshop LSDSem: Linking Models of Lexical, Sentential and Discourse-level Semantics*, Lisbon, Portugal, 2015.

Ruppenhofer, Josef and Ines Rehbein. Yes we can!? annotating english modal verbs. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012.

Modals as predictive factor for L2 proficiency level

Natalia Grabar, Bert Cappelle, Ilse Depraetere, Cyril Grandin, Benoît Leclercq

L2 language production is studied from different points of view: different linguistic phenomena (Gibbs, 1990; Moloi, 1998; Watanabe & Iwasaki, 2009; Mortelmans & Anthonissen, 2016; Mukarami et al., 2016; Ayoun et al., 2017; Romer, 2019), parallelism between L1 and L2 acquisition (Laufer & Eliasson, 1993; Chenu & Jisa, 2009; Ipek, 2009; Rabinovich et al., 2016), L1 prediction on the basis of L2 production (Jiang et al., 2014; Malmasi & Dras, 2015; Nisioi, 2015), and prediction of learner proficiency in L2 (Granfeldt & Nugues, 2007; Pilan et al., 2016; Arnold et al., 2018; Balikas, 2018). Some of these research questions have been studied manually by linguists and didacticians, while others are investigated by NLP researchers. One important question is the acquisition and use of abstract categories, such as markers of modality. For instance, in one study, the researcher analyzed the understanding of modal values (*can, could, may, might*) by English learners speaking Panjabi, an Indian language. Learners had to define the semantic value of these modals among four possible values: ability, permission, possibility, and hypothetical possibility (Gibbs, 1990). In another study, the researcher analyzed the knowledge of grammatical and modal functions of modals by children learning English (Moloi, 1998). The author noted some similarities between L1 and L2 English learners. A few papers are dedicated to the acquisition of modality in other languages, such as Japanese (Watanabe & Iwasaki, 2009) or German (Mortelmans & Anthonissen, 2016).

In our presentation, we are interested in the automatic prediction of the level of English learners on the basis of various linguistic features including modal verbs and expressions. We use the EFCamDat corpus (https://corpus.mml.cam.ac.uk/efcamdat2) built and maintained at the University of Cambridge (Geertzen et al., 2013; Huang et al., 2018). This corpus contains linguistic productions of adult learners of English from different L1 backgrounds. The exploited subset of this corpus contains 27,287 utterances with almost 2M word occurrences. These productions are categorized according to the six CECRL levels from A1 (beginners) to C2 (fluent). We use different features issued from these language productions: readability scores, n-grams of words and their comparison with the reference n-grams from the BNC and COCA corpora, modal verbs and modal expressions. We exploit different supervised learning algorithms and get up to 0.71 predictive power (F-measure performance) of fluency in English, when all the features are used. Core modals alone predict the right category for almost a third

of productions, which means that they occupy an important place in linguistic productions of English learners. We further analyze the results.

Arnold T., Ballier N., Gaillat T. & Lissón P. 2018. Predicting CEFRL levels in learner english on the basis of metrics and full texts. In *Conférence sur l'Apprentissage Automatique (CAp)*, p. 1-8.

Ayoun D. & Gilbert C. 2017. *The acquisition of modal auxiliaries in English by advanced Francophone learners*, In M. Howard & P. Leclercq, Eds., *Tense-Aspect-Modality in a Second Language: Contemporary perspectives*, p. 183-212.

Balikas G. 2018. Lexical bias in essay level prediction. In *CAp*, p. 1-5.

Chenu F. & Jisa H. 2009. Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition et interaction en langue étrangère*, 1, 1-22.

Geertzen J., Alexopoulou T. & Korhonen A. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT). In *31st Second Language Research Forum (SLRF)*.

Gibbs D. A. 1990. Second language acquisition of the English modal auxiliaries can, could, may, and might. *Applied Linguistics*, 11(3), 297-314

Granfeldt J. & Nugues P. 2007. Évaluation des stades de développement en français langue étrangère. In *TALN*, p. 1-10.

Huang Y., Murakami A., Alexopoulou T. & Korhonen A. 2018. Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1), 28-54.

Ipek H. 2009. Comparing and contrasting first and second language acquisition: Implications for language teachers. *English Language Teaching*, 2(2), 155-163.

Jiang X., Guo Y., Geertzen J., Dora Alexopoulou, L. S. & Korhonen A. 2014. Native language identification using large, longitudinal data. In *LREC*, p. 1-4

Laufer B. & Eliasson S. 1993. What causes avoidance in L2 learning. L1-L2 difference, L1-L2 similarity or L2 complexity? *Studies in Second Language Acquisition*, 15, 35-48.

Malmasi S. & Dras M. 2015. Large-scale native language identification with cross-corpus evaluation. In *Annual Conference of the North American Chapter of the ACL*, p. 1403-1409.

Moloi F. 1998. Acquisition of modal auxiliaries in English L2. *Southern African Journal of Applied Language Studies*, 6(2), 1-22.

Mortelmans T. & Anthonissen L. 2016. *German modals in second language acquisition: A constructionist approach*, In A. Stefanowitsch & T. Herbst, Eds., *Yearbook of the German Cognitive Linguistics Association*, p. 9-30

Murakami A., Michel M., Alexopoulou T. & Meurers D. 2016. Analyzing learner language in task contexts: A study case of linguistic complexity and accuracy in EFCAMDAT. In *European Second Language Association Conference*

Nisioi S. 2015. Feature analysis for native language identification. In *CICLING*, p. 1-15.

Pilan I., Volodina E. & Zesch T. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Int Conf on Computational Linguistics*, p. 2101-2111.

Rabinovich E., Nisioi S., Ordan N. & Wintner S. 2016. On the similarities between native, non-native and translated texts. *Annual Meeting of the Association for Computational Linguistics*, p. 1870-1881.

Römer U. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3), 270-292.

Watanabe S. & Iwasaki N. 2009. *The Acquisition of Japanese Modality during Study Abroad*, In B. Pizziconi & M. Kizu, Eds., *Japanese Modality*, p. 231-258.