Transitioning developmental paths in modal flavors: an experimental pilot study

Kurt Erbach & Remus Gergel, Saarland University[1]

Abstract:

This study investigates the cognitive plausibility of diachronic modal development using an experimental approach grounded in the *Human Diachronic Simulation Paradigm*. Building on historical claims that modal meanings evolve along a unidirectional path—dynamic → deontic → epistemic—we designed acceptability judgment experiments to test whether speakers accommodate hypothetical shifts in modal usage. Experiment 1 employed constructed stimuli, and Experiment 2 used naturally occurring sentences from COCA, testing three English modal expressions (*be able to*, *be allowed to*, *might*) across dynamic, deontic, and epistemic contexts. Results largely align with historical predictions (but there is also a notable exception that we discuss in more detail): *be able to* is acceptable in dynamic and deontic contexts but not epistemic; *be allowed to* remains strongly deontic and resists epistemic reinterpretation; *might* is highly acceptable in epistemic contexts and degraded elsewhere. A third experiment examined whether possibility adverbials facilitate the deontic-to-epistemic shift, revealing an interaction effect that reduces markedness in epistemic contexts. These findings support the experimental replication of diachronic tendencies while highlighting constraints on semantic change and the potential role of bridging elements. We conclude that experimental paradigms can illuminate mechanisms underlying language change and propose directions for future research integrating syntactic factors and contextual triggers.

Keywords: experimental semantics, acceptability judgment task, diachrony, modality

1 Introduction

The goal of this paper is to test the evolution of modal meanings. The background thus lies in the Labovian idea of using the present to understand the past, but with an emphasis on simulated rather than just actual current changes. The main reason to enhance the classical picture is obvious: the changes one is interested in from earlier times are quite often not practically detectable in the current environment. Recent advances in experimental semantics have set the stage for testing new observations and claims about language change (cf. Zhang et al. 2018, Fedzechkina & Roberts 2020, Fuchs et al. 2020, Gergel 2020, among others). For example, Gergel et al. (2021) suggested the Human Diachronic Simulation Paradigm to test the development paths of the English scalar additive particle *even* and the German discourse particle *doch*. This test involved testing cognates in the respective languages, *eben* and *though*, that have not undergone the same development, and other words that fulfill similar roles, like *sogar* ('even') and *nur* ('only'), but would not be expected to develop in the same way as *even* presumably did. In finding that speakers of German accept *eben* significantly more than *sogar* and *nur* in German utterances parallel to those in which *even* is found in English, Gergel et al. (2021) conclude that the paradigm shows there is convergence towards the meaning that has actually developed in the parallel language, say English, once the grammars of the German speakers are properly factored out. One limitation of this approach is the need for languages that have histories of development that are both similar and different enough.

[1] Acknowledgments: t. b. w. […]. The authors are listed in alphabetical order.

Our goal in this paper is to present a new variation of the Human Diachronic Simulation Paradigm and its usefulness in testing observations and claims about language change. In particular, both the language of study and the language of experimentation are both the same in this variation. While this does not test the robustness of historical developments across multiple languages, the present variation may potentially serve as a simpler paradigm for testing language change within a single language. Gergel, Puhl, Dampfhofer & Onea (2023) and Erbach & Gergel (2025) also simulate change within one language, too, but with non-words and outside the modal domain, while we rely in our study on lexical entries already available in the participants' grammars. An exhaustive discussion of possible approaches would lead us too far afield and we think that in general multiple approaches are needed to deepen our understanding of processes of change. For now, suffice it to say that by and large, while experiments based on artificial languages/words have the advantage that they can target rather specific readings (under the premise of successful training, not always a trivial issue), simulations based on manipulating contexts with existing natural-language items offer the advantage of taking along precisely the available meanings and the added methodological simplicity of needing no training.

Concerning our topic, it has been said that the use of particular modals follow a specific path in their evolution, namely from dynamic to deontic to epistemic, and, crucially, this is argued to be a one-directional path meaning that once a modal verb stops having deontic uses and only has epistemic uses, it will not develop deontic uses again. This pattern has been noticed across evidence from various languages including English, other Germanic languages, and beyond (for discussions see Goossens 1982, Shepherd 1982, Bybee & Pagliuga 1985, Traugott 1989, Sweetser 1990, Warner 1993, Roberts 1993, Fritz & Gloning 1997, Roberts & Roussou 2003, Gergel 2003, 2009, Nuyts, Caers & Goelen 2021, among many others). Given the breadth of such observations, we use this development path as a testing ground for our new experimental paradigm. Ultimately, the results of our experiments show the promise of this variation of the Human Diachronic Simulation Paradigm, and its potential for real-time testing of diachronic development paths.

In what follows, we first provide some background on modal development and further details on the Human Diachronic Simulation Paradigm (Section 2). In Section 3 we detail two experiments testing the modal development path, one with constructed stimuli and one with natural stimuli. In Section 4 we discuss key observations from the experiment such as the status of dynamic, deontic, and epistemic verbal modals that we tested: verbal modals *be able to*, *be allowed to*, and *might*. In Section 5, we add an experiment based on co-occurring adverbs before concluding in Section 6 that this experimental methodology can be fruitfully applied in order to replicate aspects of diachronic changes and especially to also see where key issues need to be focused more in further research.

2 Background

2.1 Modal development

In this subsection, we clarify some key developments observed in the evolution of modal meanings and the terms we use. The development of modal meanings from essentially root towards epistemic ones on the diachronic dimension has been claimed multiple times whether in descriptive traditional, functional, typological, or formal lines of linguistic work (cf. Goossens 1982, Shepherd 1982, Bybee & Pagliuga 1985, Traugott 1989, Sweetser 1990,

Warner 1993, Roberts & Roussou 2003, Gergel 2009, among many others, for discussions). The slightly more detailed schematized development is typically taken to be from dynamic to deontic and then to epistemic (and evidential) readings:

(1) Dynamic > Deontic > Epistemic.

On a typologically broader basis it is perfectly sensible to distinguish modal development paths in a more fine-grained manner, for instance, properties of evidentials from those of epistemics, but for our purposes and in the context of English we will stop at epistemics (cf. also Traugott 1989). Readings that are not epistemic (and relatedly evidential) are called root and those include dynamic and deontic ones. For convenience and understanding, we use the fairly common senses of modal flavors such as dynamic, deontic and epistemic that are not restricted to a particular framework (Palmer 1986, but see e.g. Cinque 1999 for structural deployment of very similar notions). For example, a dynamic modal typically expresses ability (sometimes physical, sometimes cognitive, etc.), a vanilla deontic modal has to do with permissions or obligations (depending on its modal force) and an epistemic modal will typically have to do with inference based on the speakers' state of knowledge. In general, it will of course also be possible to distinguish between stronger or weaker versions with respect to modal force (e.g. *must* vs. *may*, and further more fine-grained distinctions). In our own study, we will largely stay within the domain of existential (rather than universal) modal force.

The schema in (1) has been observed historically and similar observations have been made with respect to language acquisition (e.g. Lightfoot 1979, Cournane 2019). Our goal in this paper, is not to investigate the historical picture or the ample literature going together with it (cf. e.g. Eide & Gergel in prog. for an overview within Germanic), but we start out from the following assumptions:

- developments from root to epistemic are often observed and there is no reason to question them or the basis of the path established by Bybee & Pagliuca (1985) as an overall tendency;
- at the same time, when looking at the details of the developments in any language (whether English, German, Dutch, the Scandinavian Germanic languages, and similarly beyond Germanic) a number of less perfect developments arise; e.g. early epistemic that have to be explained; deontics that sometimes seem to appear later than epistemics, etc. (cf. Traugott 1989, Lühr 1997, Nuyts, Caers & Goelen 2021, among several others).

We therefore take the putatively unidirectional path claimed in the historical literature to be an interesting tendency, but which requires further clarification. Crucially, the tendency does indeed not follow directly from theoretical assumptions about the nature of modality (e.g. Kratzer 2012). We therefore follow – in spirit and on a basic level – an approach to semantic change as in Gergel, Puhl, Dampfhofer & Onea (2023) in which we take it to be worthwhile to have different developmental paths empirically checked from the perspective of their cognitive reality. This stems directly from the Labovian desire of catching language change through the prism of more directly tractable ongoing changes, but together with the realization that many interesting changes just happen not to be directly replicated around us/in present time.

The modal development path is prima facie similar to other development paths that are typically glossed under the rubric of grammaticalization and which has indeed been approached under different theoretical assumptions, but we wish to point out two things from the start. One is that grammaticalization and semantic change often interact, but on closer inspection, they do not need to (Eckardt 2006). This is important, because, rather than shifting an account of modal

meanings to another domain (e.g. syntax or prosody), we want to raise the question in a controlled way: is there a potential of confirming such one-way developments on the side of meaning? The other point is an empirical one: On closer inspection of the diachronic facts, epistemic readings *are* attested at earlier stages in multiple languages. Hence, rather than reproducing a putatively nice generalization, we wish to test – with an open mind and on the basis of a new methodological approach – whether the claimed development can hold scrutiny when a partially similar set-up is constructed and tested as the ones presumably available historically. We cannot emphasize the word 'partially' strongly enough, since no experimental replication will be an exact depiction of a historical change. But we do move beyond so-called artificial language paradigms and seek to implant a relatively natural language environment into our otherwise simple experimental design (while also discussing some more involved alternatives).

2.2 The Human Diachronic Simulation Paradigm

The Human Diachronic Simulation Paradigm is an empirical method designed, as the name states, to simulate the native-speaker judgments one would seek when testing what sorts of diachronic change might occur in a given language (Gergel et al. 2021). As formal semantics largely relies on speaker intuition to distinguish well formed structures from infelicitous or ungrammatical ones (Chierchia and McConnel-Ginet, 2000: 2-3), diachronic semantics could be seen to be at a disadvantage because it lacks access to native speakers of the previous form of the language. Gergel et al. (2021) argue that by testing a historical development path with speakers of a related language or dialect (in which certain semantic premises hold) that have yet to realize a specific change, their judgments can be used in place of speakers of the previous form of the language. In particular, the idea is that these speakers will accept use of their language that exemplifies the change that has taken place elsewhere historically compared to use of their language that exemplifies unattested changes. In this way, acceptability judgements can be used to simulate the conditions under which diachronic change has occurred.

One means of simulating language change to test hypotheses on the topic has been to use artificial language learning experiments (Roberts and Snelle, 2020: 194). Artificial language learning experiments have been characterized as those in which the experimenter constructs a portion of one or more constructed languages that differ minimally from each other so as to precisely control the input that learners are exposed to (Culbertson and Schuler, 2019: 354). Constructed languages are those consciously created by individuals, often for the purpose of philosophical thought experiments, linguistic experiments, or fictional peoples in literature (Goodall, 2023: 419). For example, Kirby, Cornish and Smith (2008) use iterated learning of an artificial language by participants in an experiment to show that patterns and structures emerge without any intention by the participants or experimenters. Roberts and Fedzechkina (2018) also use iterated learning of an artificial language to illustrate how social biases may modulate the loss of redundant forms across generations. In the present paper, we follow an approach fully embedded in natural language, that is, while certain stimuli in our experiments can be called artificial in that they do not exist in the English that is known to be spoken anywhere at this time, we do not make use of a constructed language like those in the experiments mentioned above.

We take as inspiration Gergel et al. (2021), who use the Human Diachronic Simulation Paradigm to test the development of the English discourse particle *even* and the German discourse particle *doch*, which is said to retrieve a proposition p from the common ground and signal a contrast (Repp 2013). In both experiments, the developmental path of the respective

particles was tested in the other language, where the development has not taken place: *eben* ('even') does not function as a discourse particle the way *even* is an additive of improbability, and *though* does not function as a discourse particle the way *doch* does. Within the structure of the Human Diachronic Simulation Paradigm, the development of *even* from a status similar to *eben*, as in an *even surface*, was tested by using *eben* in German sentences matching the use of *even* as a discourse particle in English (2).

(2) Eben Maria, die sonst immer zuhause beliebt, ist gekommen.
even Maria she otherwise always at.home stays, is come
"Even Maria, who otherwise always stays home, is coming."

The stimuli consisted of sentences where different focus-sensitive particles *sogar* ('even'), *nur* ('only'), and *auch* ('too/also') naturally occur, but were replaced with *eben* ('even'). The stimuli were tested for acceptability on a 7-point Likert scale from *völlig akzeptabel* ('fully acceptable') to *gar nicht akzeptabel* ('not at all acceptable'). The acceptability of *eben*, when used in place of *sogar*, was more acceptable ($p<0.01$) than when used in place of *nur* and *auch*. The higher acceptability of *eben* in place of *sogar* rather than other competitors is taken to indicate that the discourse particle meaning of *sogar* is calculable from the meaning of *eben*. Similar results were obtained from the experiment with *though* to test the development of *doch*. In summary, the Human Diachronic Simulation Paradigm was developed as a means of looking at diachrony with synchronic empirical work. In the next section we describe how we used this paradigm albeit entirely within one language, namely English.

3 Testing the modal-development path

This experiment tests three different modal lexemes across three different contexts of use, thereby testing theories of how the use of a given modal lexeme changes over time. We designed an experiment to aggregate acceptability judgments of specific modal expressions across utterances with dynamic, deontic, or epistemic meaning. We specifically capitalized on modal expressions that are empirically not considered to have the vanilla ambiguity of modal flavors assumed by semantic theory, but in fact are standardly taken to have one prevalent modal flavor in actual use. For example, each of the constructed sentences in (3)-(5) exemplify dynamic (ability/circumstances), deontic (permission/duty), and epistemic (knowledge) modality respectively in the verbal modals *be able to*, *be allowed to*, and *might*.

(3) Because Sara's car is parked in her driveway, she is able to easily load it with the luggage.
(4) Since Anne's truck is parked on her property, she is allowed to leave it there indefinitely.
(5) Given Maureen's vehicle is parked in front of her apartment, she might be home.

3.1 Experiment 1: Constructed sentences

The sentences in (3)-(5) were taken as "original" items and used to create "modified" items, where each targeted verbal modal was swapped across contexts. For example, (4) is taken as a deontic context and the dynamic modal *is able to* is put into this context yielding the sentence in (6). All modals were put into all contexts thereby testing the acceptability of modals both in contexts that correspond to the anticipated development (forwards along the evolutionary path) and in contexts that go against the anticipated development (backwards along the evolutionary path). Given existing theory, we anticipate (6)-(8) to receive higher acceptability judgments,

as they exemplify possible future development of BE *able to* and BE *allowed to*, in comparison to items like (9)-(11), which do not fit with the assumed evolution path of deontic and epistemic modality (i.e. they are backwards steps on the path).

(6) Since Anne's truck is parked on her property, she is able to leave it there indefinitely.
(7) Given Maureen's vehicle is parked in front of her apartment, she is able to be home.
(8) Given Maureen's vehicle is parked in front of her apartment, she is allowed to be home.
(9) Because Sara's car is parked in her driveway, she is allowed to easily load it with the luggage.
(10) Because Sara's car is parked in her driveway, she might easily load it with the luggage.
(11) Since Anne's truck is parked on her property, she might leave it there indefinitely.

In summary, we have a 3x3 Latin Square design where the condition MODAL has three levels (be_able_to, be_allowed_to, and might) and the condition CONTEXT has three levels (dynamic, deontic, and epistemic). To avoid possible influence of same co-text—e.g. our original epistemic sentence occurring with different modals (3), (4), and (5)—each co-text constructed from the same original item were spread across lists, each given to a different group of participants, so there were three lists, each consisting of 36 target items and 36 filler items. Four practice items were given, two with instructions (one clearly acceptable and one clearly unacceptable) and two without (one acceptable and one unacceptable), The test items (target and filler) in each list were presented in randomized order. Items were randomized via the randomization built into the survey hosting platform used, PCIbex (Zehr and Schwarz 2018). Apart from the two practice items with detailed instructions, each item was presented underneath the same question prompt, "Read the sentence and rate how natural the word use sounds to you", and underneath the test item, participants will provide their response via a sliding scale labeled with "completely unnatural" on the left side of the scale and "completely natural" on the right. The slider first appeared in the middle and it was moved by clicking and dragging towards whichever side seems most appropriate to the participant. In other words, the measured variable is sentence naturalness which is measured via participant response by moving a slider on a scale. Though the participant did not see the underlying values of the scale, it is from 0 to 999.

In an attempt to prevent rigidly binary use of the sliding scale for unfamiliar constructions, participants were told the following:

> Some of these sentences come from a dialect of English that is different from yours. This dialect has been undergoing some changes in how certain words are used. We're curious what English speakers like yourself think of these changes compared to standard sentences. So, we want you to rate how natural the formulation of each sentence is to you.

This deliberate deception in which we tell participants some of the data is from a particular dialect is in line with Human Dialect Simulation Paradigm, and is designed to direct the participants to be more accommodating. The direct acknowledgement of a standard is to also direct the participants to not simply accept all of the given forms. These behaviors cannot be guaranteed, of course, so at the very least these instructions give the task a bit more context.

Participants were recruited via Prolific and filtered for being born and raised in the US speaking only English until 18, currently using English as their primary language, and not having participated in one of the other iterations of this experiment (one of the other lists). Participants

were able to quit at any time, and only if they completed the study were they compensated (a rate of €12.41 per hour). 20 participants were recruited per list (60 total) based on a pilot study run with 98.17% power, registered on OSF[2].

We predicted that there would be an effect of MODAL, CONTEXT, and MODAL*CONTEXT interaction given that BE *able to* should be relatively acceptable in all contexts, BE *allowed to* should be relatively acceptable in two contexts (deontic and epistemic), and *might* should only be acceptable in one context (epistemic).

The participants' judgments of the experimental items are summarized in Table 1 and Figure 1. Four participants' judgments of original sentences averaged less than 550 (out of 999) and were therefore excluded. Looking at the averages alone, we see that the judgements generally follow the predictions for acceptability given the modal development path: BE *able to* is best in dynamic contexts (its current use), is degraded in deontic contexts (its next step in development), and is further degraded in epistemic uses (its subsequent step in development); BE *allowed to* is best in deontic contexts (its current use), is degraded in both dynamic and epistemic contexts (its previous and future uses according to the development path); and *might* is best in epistemic contexts (its current use), is degraded in deontic contexts (its previous stage of use), and is further degraded in dynamic contexts (its use two stages previous to now).

modifier	context	mean	sd
be_able_to	dynamic	807.0085	264.2099
be_able_to	deontic	777.7773	271.4531
be_able_to	epistemic	583.1234	347.7521
be_allowed_to	dynamic	578.6287	353.8097
be_allowed_to	deontic	807.1181	264.9263
be_allowed_to	epistemic	555.0932	344.0943
might	dynamic	684.4746	321.2134
might	deontic	752.9234	288.8427
might	epistemic	821.8051	229.7820

Table 1: Average rating and standard deviation of sentence judgment

[2] https://osf.io/4qzya/

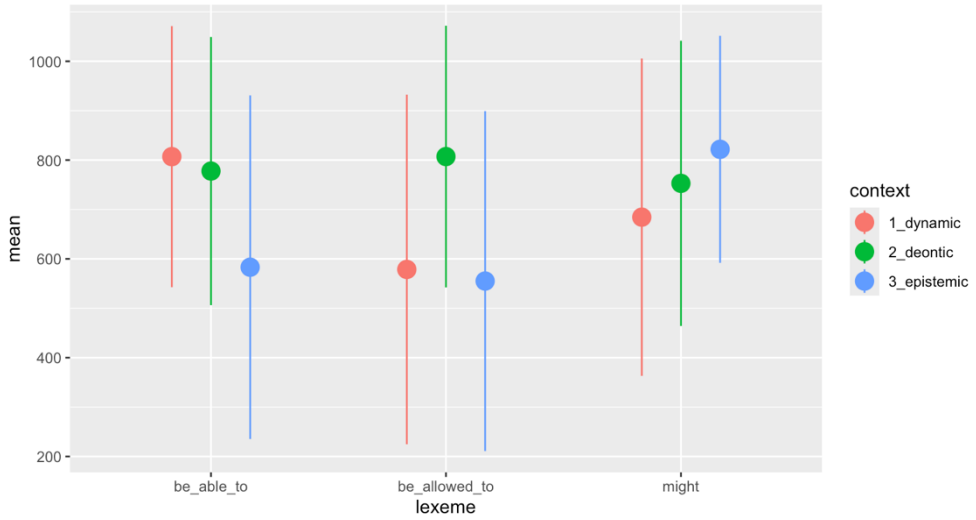

Figure 1: Mean acceptability ratings, 0 (worst) to 1000 (best), for Experiment 1.

We modeled the results in R (R Core Team 2021) using the lme4 package for linear mixed-effects models (Bates and Maechler 2009). The model is given in (12) and is summarized in Table 2. As expected, we see no evidence of a difference between original sentences, that is that be_able_to:dynamic, be_allowed_to:deontic, and might:epistemic are all be judged to be similarly acceptable. The average rating of each differs by a maximum of about 14 on the 1000-point scale, and this is less than the difference between be_able_to:dynamic and be_able_to:deontic which was not found to be significant (SE = -25.44, p = 0.56163). In other words, we might assume that the original sentences—i.e. the modals used in their normal context of use—are acceptable as a baseline of comparison for modified sentences testing the modal development path.

(12) lmer(Response ~ modifier*context + (1|Participant)+ (1|item),REML=FALSE)

Fixed effects	Estimate	SE	df	t value	Pr(>|t|)
(Intercept)	802.37	35.89	153.26	22.359	< 2e-16
be_allowed_to	-226.23	43.71	106.52	-5.176	1.08e-06
might	-121.05	43.72	106.66	-2.769	0.00664
deontic	-25.44	43.69	106.40	-0.582	0.56163
epistemic	-223.16	43.74	106.81	-5.102	1.47e-06
be_allowed_to:deontic	253.48	61.79	106.39	4.102	8.04e-05
might:deontic	91.50	61.82	106.61	1.480	0.14180
be_allowed_to:epistemic	198.95	61.83	106.65	3.218	0.00171
might:epistemic	358.68	61.84	106.75	5.800	6.85e-08
Random effects					
Groups	Variance	Std.Dev.			

item	8254	90.85
Participant	19912	141.11
Residual	63177	251.35

Table 2: Summary of linear model (12) results for Experiment 1.

Starting with the development of BE *able to*, as said in the previous paragraph, the difference between be_able_to:dynamic and be_able_to:deontic which was not found to be significant (SE = 43.69, p = 0.56163). In other words, no evidence was found that using this modal in the next context of its development path significantly degrades acceptability. One might take this to mean that the deontic contexts were flexible enough to allow dynamic modals and dynamic meaning, or that BE *able to* is already well on its way to being used as a deontic modal. In either case, BE *able to* is definitely unacceptable in epistemic contexts as this combination caused a significant degradation in acceptability (SE = 43.74, $p < 0.001$).

The results of our experiment suggest that BE *allowed to* is firmly in its place as a deontic modal, because judgments of this modal in dynamic and epistemic contexts are significantly worse. The difference between BE *allowed to* in its original context, deontic, and its previous stage in the development path, dynamic, is greater than that between be_able_to:dynamic and be_able_to:epistemic, which was found to be a significant difference. The difference between be_allowed_to:deontic and be_allowed_to:epistemic is also significant in this way. In summary, BE *allowed to* was rated significantly worse in dynamic and epistemic contexts than in deontic ones.

Lastly, the judgments of sentences containing *might* appear at first glance to directly follow as predicted by the modal-evolution path: sentences in *might*'s current stage, epistemic, are rated highest, sentences one stage previous, deontic, are rated lower, and sentences two stages previous, dynamic, are rated lower still. The difference between the distance from the intercept to might:dynamic versus that to might:denotic is relatively small and not significant (SE 61.82, p = 0.1418). The same could be said of the difference between the distance from the intercept to might:deontic versus that to might:epistemic, given this is an even smaller difference. On the other hand, difference between the distance from the intercept to might:dynamic and that to might:epistemic is significant (SE = 61.84, $p < 6.85e-08$). In other words, the acceptability of *might* in deontic contexts is relatively similar to that of *might* in epistemic contexts, but different from that of *might* in dynamic contexts. Altogether, this reflects the modal-evolution path in that use of *might* in its previous stage is not marked while use of *might* in two stages previous is marked.

We conducted an ANOVA of the linear model to test for effects of modifier and context alone, and the results are summarized in Table 3. This test shows that, overall, there are effects of modifier, context, and modifier:context interaction ($p < 0.001$). In other words, as predicted, overall each modal is judged differently from the others, and each context sees different judgments as well. Moreover, the judgment of acceptability of each modal is modulated through the context. These results would not be expected if each modal could simply be used more or less to the same degree of acceptability with any meaning in its development path. Instead, what we see is that participants rate the modals differently depending on the contexts in which they occur as would be expected given the development path.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
modifier	1152889	576444	2	106.59	9.1243	0.0002199
context	1685366	842683	2	106.64	13.3385	6.758e-06
modifier:context	3455033	863758	4	106.60	13.6721	4.951e-09

Table 3: Type III Analysis of Variance Table with Satterthwaite's method

Summarizing, the results of the linear model of this experiment relative to our predictions, the development path is illustrated with our results: judgments of BE *able to* decrease the further the modal is used ahead of its current place in the development path; judgments of BE *allowed to* are best in its current stage of use (deontic) and significantly worse in prior and future stages (dynamic and epistemic respectively); and judgments of *might* are best in its current stage and increasingly worse the further back along the development path it is used (first back to deontic, then further back to dynamic).

3.2 Experiment 2: Naturally occurring sentences

While the results of the first experiment generally fit the development path, there are some specific results that do not fit the predictions of the path and the Human Diachronic Simulation Paradigm. In particular, we might have expected to find a significant difference between the acceptability of BE *able to* in dynamic contexts (its current use) and in deontic contexts (its next stage in the development path), which we did not find. Additionally, we also would have expected that BE *allowed to* is significantly more acceptable in epistemic contexts (its next stage in the development path) than in dynamic contexts (its previous stage in the development path), and that the difference between *might* in deontic contexts (its previous stage in the development path) is significantly less acceptable than in epistemic contexts (its current stage in the development path). While these results are not necessarily problematic, they do present at least the following question: could the constructed stimuli have given rise to these differences between participant judgments and our expectations? In other words, one reason for these differences between predictions and results may be due to the artificial, constructed sentences, while naturally occurring sentences may not necessarily yield the same results.

To investigate the extent to which our constructed sentences in particular gave rise to these results as opposed to the course of modal development, we collected utterances containing dynamic uses of BE *able to*, deontic uses of BE *allowed to*, and epistemic uses of *might*, from the Corpus of Contemporary American English (COCA) (Davies 2008). We edited the sentences as little as possible so they would fit the experiment in a comprehensible way without a larger context. As before, we then created a set of modified sentences from these original sentences by replacing the modal in each with the other modals being tested. Again, each original sentence and the modified versions made from it was distributed across a different list, and participants were recruited via Prolific and filtered for being born and raised in the US, speaking only English until 18, currently using English as their primary language, and not having participated in one of the other iterations of this experiment (one of the other lists). Participants were able to quit at any time, and only if they completed the study were they compensated (a rate of €12.41 per hour). 20 participants will be recruited per list (60 total) based on the results of a the pilot experiment being 98.17% power.

The results are basically the same as it was with our constructed sentences for our dynamic contexts and deontic contexts, but it it is different for epistemic contexts. The results are summarized in Table 4 and Figure 2 and the results of the linear model (12) are summarized in Table 5.

modifier	context	mean	sd
be_able_to	dynamic	695.7069	308.5681
be_able_to	deontic	711.7650	284.2241
be_able_to	epistemic	482.7382	337.6426
be_allowed_to	dynamic	589.2650	341.3481
be_allowed_to	deontic	750.7931	261.2978
be_allowed_to	epistemic	514.4292	337.6399
might	dynamic	619.6840	335.7339
might	deontic	632.2900	322.1365
might	epistemic	750.1552	289.7997

Table 4: Average rating and standard deviation of sentence judgment.

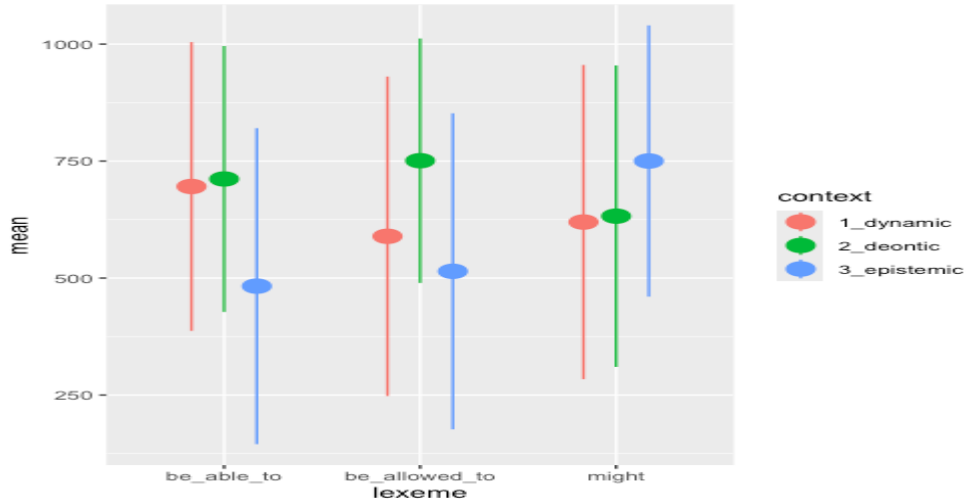

Figure 2: Mean acceptability ratings, 0 (worst) to 1000 (best), for Experiment 2.

Fixed effects	Estimate	SE	df	t value	Pr(>|t|)
(Intercept)	694.036	37.175	146.958	18.669	< 2e-16
be_allowed_to	-104.684	46.598	106.069	-2.247	0.0267

might	-75.626	46.642	106.466	-1.621	0.1079
deontic	17.116	46.601	106.087	0.367	0.7141
epistemic	-213.516	46.614	106.207	-4.581	1.27e-05
be_allowed_to:deontic	141.786	65.904	106.090	2.151	0.0337
might:deontic	-5.296	65.944	106.341	-0.080	0.9361
be_allowed_to:epistemic	137.965	65.899	106.068	2.094	0.0387
might:epistemic	343.469	65.941	106.328	5.209	9.37e-07
Random effects					
Groups	Variance	Std.Dev.			
item	9322	96.55			
Participant	17391	131.87			
Residual	71881	268.11			

Table 5: Summary of linear model (12) results for Experiment 2.

As before, BE *able to* is currently a dynamic modal and participants do not appear to rate it differently in deontic contexts (Est = 17.116, SE = 46.601, p = 0.7141), while they rate it markedly lower in epistemic contexts (Est = -213.516, SE = 46.614, p < 0.001). The estimate of be_allowed_to:deontic (Est = 141.786, SE = 65.904, p < 0.05) shows that there is a significant difference between participant ratings of BE *allowed to* in deontic contexts over and above the differences from the intercept, that is, BE *allowed to* in dynamic contexts. From this, one might extrapolate that BE *allowed to* is also significantly worse in epistemic contexts than deontic contexts. However, the difference between BE *allowed to* in dynamic versus epistemic contexts is likely not significant given the difference between be_able_to:dynamic and might:dynamic is similar and not significant (Est = -75.626, SE = 46.642, p = 0.1079). The difference between might:epistemic and might:deontic is likewise large enough to be significantly different given it is larger (750-632=118) than the difference between be_able_to:deontic and be_allowed_to:deontic (Est. = 104.684, SE = 46.598, P < 0.05). However, the difference between might:dynmaic and might:deontic (632-619=13) is likely not significant. These results align with those of Experiment 1, in which sentences were generally rated the same way.

It might be notable that, *might* is not markedly worse in dynamic contexts than BE *able to* (Est. = -75.626, SE = 46.642, p = 0.1079), though given the differences found in previous studies with constructed stimuli, we assume this is a fluke of the somewhat lower-rated original sentences in experiment 2. For example the average rating of be_able_to:dyanmic in Experiment 2 is 695 while it is 807 in Experiment 1. At the same time the average rating of might:epistemic in Experiment 2 is 750 compared to 821 in Experiment 1. So, while the overall patterns align across the constructed sentences in Experiment 1 and the naturally occurring sentences in Experiment 2, there are some differences that might be due to the particular naturally occurring sentences that were selected for this experiment.

With respect to the evolution of these modals, these experiments could be taken to show several things about BE *able to*, BE *allowed to*, and *might* and their trajectory along the evolution path of modals (dynamic > deontic > epistemic). First, BE *able to* seems to be fully acceptable in dynamic and deontic contexts ($p=0.714$), though not at all in epistemic ($p<0.001$). Second, BE *allowed to* is firmly deontic, and while it is markedly worse in both dynamic and epistemic contexts ($p<0.05$), and no difference between ratings of BE *allowed to* in dynamic and epistemic contexts was found. Lastly, *might* is equally ($p=0.936$) markedly worse in both deontic and dynamic contexts than epistemic ($p<0.001$). Altogether, this experimental paradigm shows the path of development and the positions of BE *able to*, BE allowed to, and *might* on that path.

4 Discussion

We can distill five key observations from our investigation contrasting speaker reactions in the simple set-up of our experiments and the diachronic line of research. Four of these observations confirm expectations generated by historical development, but one crucially does not. We present them in turn and focus on the diverging outcome and why we think it is still informative rather than just puzzling.

(A) *might* is highly degraded when placed in non-epistemic contexts (both deontic and dynamic).

This (package of factually two observations) is expected based on the assumption of unidirectionality. Epistemics don't go back on the path of change. Notice also from Experiment 2 above that dynamics are even worse than deontics for *might.*

(B) *be able to* does not easily fit epistemic contexts.

The reason why (B) may be expected is that a dynamic modal would become directly epistemic and a crucial intermediate step is skipped in this type of simulated development. (Why the intermediate step from deontic to epistemic is already difficult enough to achieve, we discuss under (E) below.)

(C) *be able to* is very easily adaptable to deontic contexts.

This is in line with the general diachronic trajectory and might strengthen the earlier observation, namely that the problem why *be able to* does not easily become epistemic in our judgment tasks can at least not be blamed on the lack of the "first" step (from the original (1)), but that it must be related to something else.

(D) *be allowed to* does not easily become dynamic.

This result is expected in a similar way as (A). We take it that from a historic-developmental it would be argued to be as an instance of trying to go against a largely unidirectional developmental tendency. (C) and (D) moreover show an expected asymmetry.

(E) *be allowed to* does not become epistemic.

(E) is unexpected at first blush clearly a strong blow against the core of the grammaticalization and semantic change line of work. But it is worth looking at the details given it may point to how such changes may (not) occur.

Recall that we avoided modals like *can, may* etc. in our selection since they are ambiguous already synchronically (and barely anything would have to be 'enforced' or tested in a simulation). The choice of an unambiguous semi-modal expression like *be allowed to* (and we take it that others such as *be required/forced/… to* would behave quite similarly) is that we have two components of meaning that might not have been so prominent in actual changes. One is that we have an authority clearly visible in such expression that we take to crucially stand in the way of an epistemic reinterpretation. That is, only after this component of meaning has been obliterated, should an epistemic reinterpretation be possible. The other possible explanation is that the agentivity of the subject is required under such modal expressions like *be allowed to*, as already pointed out by Bybee and Pagliuga (1985), and epistemics do not have such a requirement. We are not certain whether this is a genuine explanation, since epistemics might not *require* agentivity, but they do allow it (cf. *Jane must be riding her bike very carefully*). But it is true that agentivity is available in the deontic *be able to* and distinguishing between the two possibilities might require more research.

To test whether adverbials may facilitate the use of deontic modals in epistemic contexts, we used a 2x2 Latin-square design with the factors modal FLAVOR and MODIFICATION with the respective levels DEONTIC, EPISTEMIC and ADVERBIAL, NONE. For example, a DEONTIC:AVERBIAL stimulus would look like (13), where the adverb *possibly* to modifies BE *allowed to* in a deontic context. This would be tested in addition to the same modified modal in an epistemic context as is seen with the EPISTEMIC:ADVERBIAL example in (14), where it is presumably marked, but relatively less-marked than the unmodified EPISTEMIC:NONE example in (15). Finally, (16), DEONTIC:NONE, completes the paradigm.

(13) Since Anne's truck is parked on her property, she is possibly allowed to leave it there indefinitely.
(14) Given Maureen's vehicle is parked in front of her apartment, she is possibly allowed to be home.
(15) Given Maureen's vehicle is parked in front of her apartment, she is allowed to be home.
(16) Since Anne's truck is parked on her property, she is allowed to leave it there indefinitely.

Testing this paradigm as before, we would expect the same high acceptability of DEONTIC:NONE that we saw with be_able_to:DEONTIC in the previous experiments, and the same markedness of EPISTEMIC:NONE as be_allowed_to:EPISTEMIC as in the preivous experiments because these items are the same. However, in this third experiment, if adverbials like *possibly* facilitate the development of deontic modals into epistemic ones, then we should expect to see EPISTEMIC:ADVERBIAL is less marked than EPISTEMIC:NONE. Moreover, if there is indeed an effect of adverbials on the acceptability of deontic modals like BE *allowed to* in epistemic contexts, then we should also expect to see an interaction between the two.

5 The effect of possibility adverbials on deontic modals

Using the experimental setup described above, we tested the effect of possibility adverbials on deontic modals. We split 12 paradigms resembling (13)-(16) across four lists with an equal number of filler items. We hired 20 participants per list via Prolific and paid them at a rate of €12.41 per hour, and they were able to quit participating at any time without penalty and without payment. Participants were given the same instructions detailed above about analyzing a set of sentences, some of which come from a dialect different from theirs. Participants likewise were given the same set of four training items to be rated by moving an indicator on a sliding scale between "completely unnatural" on the left side of the scale and "completely natural" on the right. The scale had the same underlying values from 0 to 999 that were unknown to the participants. Attention was controlled with four questions of similar design but clearer answers with respect to naturalness, and participants' answers were thrown out if they answered more than 75% of the control questions incorrectly—i.e with a score on the wrong side of the scale (e.g. ≤500 when the answer was clearly "completely natural"). Four participants' answers met this criteria and were therefore thrown out.

The participants' judgments of the experimental items are summarized in Table 6 and Figure 3. Four participants judgments for the control items met the rejection criteria and were therefore excluded. Looking at the averages alone, one might be tempted to conclude that adverbs do not provide a mitigating effect given EPISTEMIC:ADVERBIAL has the lowest judgments of all conditions. However, the fact that DEONTIC:ADVERBIAL is also lower than DEONTIC:NONE, and the fact that the distance between DEONTIC:ADVERBIAL and EPISTEMIC:ADVERBIAL appears to be much less than that between DEONTIC:NONE and EPISTEMIC:NONE both give reason for pause. This smaller distance may be the sort of effect we are looking for, namely one that makes the difference between DEONTIC and EPISTEMIC uses of BE *allowed to* smaller than is seen in their unmodified counterparts.

flavor	modification	mean	sd
deontic	adverbial	653.7692	295.7525
episetmic	adverbial	507.9316	329.7106
deontic	none	855.2112	232.0591
episetmic	none	568.6052	341.7492

Table 6: Average rating and standard deviation of sentence judgment in Experiment 3

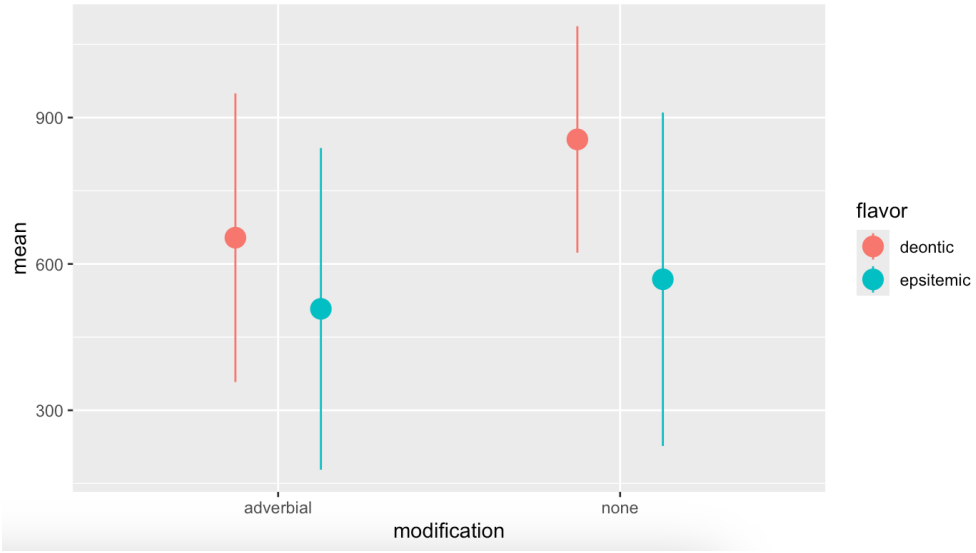

Figure 3: Mean acceptability ratings, 0 (worst) to 1000 (best) and sd, for Experiment 3.

We modeled the results in R (R Core Team 2021) using the lme4 package for linear mixed-effects models (Bates and Maechler 2009). The model is given in (17) and is summarized in Table 7. The key here is that we see a significant effect between all differences, suggesting that the acceptability of BE *allowed to* in deontic and epistemic contexts is modulated through modification. First and unsurprisingly, we see a significant difference ($p<0.01$) between DEONTIC:ADVERBIAL and EPISTEMIC:ADVERBIAL; this pattern is in line with the first experiment where use of BE *allowed to* is marked in the epistemic contexts. Second, and to our surprise, there is also a significant difference between DEONTIC:ADVERBIAL and DEONTIC:NONE ($p<0.001$). Because there is no theoretical basis for this difference—i.e. because possibility adverbials can occur with deontic modals—one might assume that this difference is due to processing cost: inverse correlations are often found between sentence complexity and acceptability judgments (Meyers 2017). Third, and in line with out predictions, there is a significant difference between DEONTIC:ADVERBIAL and EPISTEMIC:NONE ($p<0.05$) on top of the aforementioned differences. Again this suggests there is an interaction between modal flavor and modification when it comes to the acceptability of BE *allowed to*.

(17) lmer(Response ~ flavor*modfication+ (1|Participant)+ (1|item),REML=FALSE)

Fixed effects	Estimate	SE	df	t value	Pr(>|t|)
(Intercept)	654.19	36.70	65.75	17.825	< 2e-16
epistemic	-145.83	46.47	43.70	3.138	0.00304
none	203.65	46.50	43.81	4.380	7.29e-05
epistemic:none	143.65	65.74	43.77	-2.185	0.03429
Random effects					
Groups	Variance	Std.Dev.			

item	9866	144.36
Participant	20840	99.33
Residual	60232	245.42

Table 7: Summary of linear model (17) results for Experiment 3.

To test whether this perceived interaction effect is significant, we conducted an ANOVA between the linear model in (17), which includes interaction, and one that does not. The results are summarized in Table 8. This test shows that there is a significant difference between the two models—i.e. that there is a significant difference when the model includes an interaction between modal flavor and modification. This can be taken to suggest that modification does have an effect on the acceptability of BE *allowed to* in epistemic contexts. Moreover, given the difference between DEONTIC:ADVERBIAL and EPISTEMIC:ADVERBIAL is smaller than the difference between DEONTIC:NONE and EPISTEMIC:NONE, one might conclude that possibility adverbials do indeed lower the relative markedness BE *allowed to* in epistemic contexts.

	npar	AIC	BIC	logLik	-2*log(L)	Chisq	Df	Pr(>Chisq)
no interaction	6	13125	13154	-6556.7	131113			
interaction	7	13123	13157	-6554.4	13109	4.536	1	0.03319

Table 8: Analysis of variance for effect of interaction in Experiment 3

Summarizing, the results of the linear model and the ANOVA thereof for Experiment 3 can be taken as support for the idea that possibility adverbials help to facilitate the evolution of deontic modals into epistemic ones. When no adverbials are present, the difference between BE *allowed to* in deontic and epistemic contexts is greater than when adverbials are present.

6 Conclusion

We have shown that the methodology of conducting experiments in order to replicate diachronic changes can be fruitfully applied to the case of modals. Some of our results clearly support the claims of the historical literature, but we also pointed out a key result that does not appear as directly consonant with the diachronic development at first glance. A pure modal expression does not easily transfer from a deontic to an epistemic meaning even when the contextual factors are right. There are three key takeaways from this from a developmental perspective. One is that the syntax, which was outside of our current investigation which focused on the semantics of the modals, might have played a larger role in the actual changes. Integrating syntactic change under quick lab conditions is not trivial, but certainly a task that can be approached in future research (cf. Gergel et al. 2023, Eide & Gergel 2025). Second, it is perfectly possible that beyond the general contextual factors being right, a key bridging factor might have been needed. This could be epistemic adverbs that color the originally deontic modal epistemic after reinterpretation. This specific step is relatively straightforward to achieve in future research. Third, a more critical note can also emerge. One possibility why replicating the change deontic → epistemic appears comparatively harder than other is that the change either required very specific conditions or that the actual modals that underwent the

change were already more predisposed – and possibly already potentially epistemic – before the observed changes in the historical records took place. This question links up to the previous two pointed out in this conclusion and opens up a further research paradigm.

References

Bates, D. and M. Maechler (2009). lme4: Linear mixed-effects models using s4 classes. r package version 0.999375-31. available at http://cran.r-project.org/package_lme4.

Bybee, Joan L., and William Pagliuca. 1985. Cross-linguistic comparison and the development of grammatical meaning. *Historical semantics, historical word formation* 59, 67-71.

Chierchia, Gennaro and McConnel-Ginet, Sally. 2000. *Meaning and Grammar.* MIT Press.

Cinque, Guglielmo. 1999. *Adverbs and Functional Heads: A cross-linguistic perspective*. Oxford University Press.

Cournane, Ailís. 2019. A developmental view on incrementation in language change. *Theoretical Linguistics* 45: 127-150.

Culbertson, Jennifer and Kathryn Schuler. 2019. Artificial language learning in children. *Annual Review of linguistics* 5: 353-373 DOI: 10.1146/annurev-linguistics-011718-012329

Davies, Mark. 2008. The corpus of contemporary American English (COCA). https://www.english-corpora.org/coca/.

Eckardt, Regine. 2006. *Meaning Change in Grammaticalization: An enquiry into semantic reanalysis*. Oxford University Press.

Eide, K. M. & R. Gergel. 2025. Germanic modal trajectories: meaning and structure. Ms. NTNU & Saarland University.

Erbach, Kurt and Remus Gergel. 2025. Testing dialects with simulations: the status of pseudo-partitives in US English. *Linguistica Brunensia* 73 (1): 93-111. DOI: 10.5817/LB2025-38532

Fedzechkina, Masha & Gareth Roberts. 2020. Learners sacrifice robust communication as a result of a social bias. https://doi.org/10.31219/osf.io/usfhz.

Fritz, Gerd & Thomas Gloning. 1997. *Untersuchungen zur semantischen Entwicklungsgeschichte der Modalverben im Deutschen*, Berlin: Max Niemeyer Verlag.

fritz

Fuchs, Martín, Ashwini Deo & María Mercedes Piñango. 2020. The Progressive-to-Imperfective shift. Contextually determined variation in Rioplatense, Iberian, and Mexican Altiplano Spanish. In Alfonso Morales-Front, Michael J. Ferreira, Ronald P. Leow and Cristina Sanz (eds.), *Hispanic Linguistics: Current issues and new directions*, 119–136. Amsterdam: Benjamins.

Gergel, Remus. 2003. Modal syntax: Detecting its parameters with VP-ellipsis. *SKY Journal of Linguistics* 16: 27-56.

Gergel, Remus. 2009. *Modality and Ellipsis: Diachronic and Synchronic Evidence*. Mouton de Gruyter.

Gergel, Remus. 2017. Dimensions of variation in Old English modals. In M.L. Rivero, A. Arregui, A. Salanova, (eds.), *Modality across Syntactic Categories.* 179-207. Oxford University Press.

Gergel, Remus. 2020. Sich ausgehen: Actuality entailments and further notes from the perspective of an Austrian German motion verb construction. *Linguistic Society of America*, *5*(2). 5-15. https://doi.org/10.3765/plsa.v5i2.4790.

Gergel, Remus, Martin Kopf-Giammanco, and Maike Puhl. 2021. Simulating semantic change: a methodological note. *Experiments in Linguistic Meaning* 1, 184–196.

Gergel, Remus, Maike Puhl, Simon Dampfhofer, and Edgar Onea. 2023. The rise and particularly fall of presuppositions: Evidence from duality in universals. *Experiments in Linguistic Meaning, 2,* 72-82.

Goodall, Grant. 2023. Constructed Languages. *Annual Review of Linguistics* 9, 419--437. DOI: 10.1146/annurev-linguistics-030421-064707

Goossens, Louis. 1982. On the development of the modals and the epistemic function in English. In *Papers from the Fifth International Conference on Historical Linguistics*, ed. Anders Ahlqvist, 74-97. John Benjamins.

Kirby, Simon, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS* 105 (31), 10681-10686. DOI: 10.1073/pnas.0707835105

Kratzer, Angelika. 2012. *Modals and Conditionals*: New and revised perspectives. Oxford University Press.

Lightfoot, David. 1979. *Principles of Diachronic Syntax*. Cambridge University Press.

Lühr, Rosemarie. 1997. Zur Semantik der althochdeutschen Modalverben. In *Untersuchungen zur semantischen Entwicklungsgeschichte der Modalverben im Deutschen*, Fritz, Gerd & Gloning, Thomas (eds.): pp. 159-175. Niemeyer.

Myers, James. 2017. Acceptability judgments. In *Oxford research encyclopedia of linguistics*. DOI: 10.1093/acrefore/9780199384655.013.333

Nuyts, Jan, Wim Caers & Henri-Joseph Goelenk. 2021. System and variation in the Dutch modals. In *Cognitive Sociolinguistics Revisited*, eds. Kristiansen, G., Franco, K., De Pascale, S., Rosseel, L., & Zhang, W., pp. 242-250. DeGruyter.

Palmer, F.R. 1986. *Mood and Modality*. Cambridge University Press.

R Core Team (2021). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Repp, Sophie. 2013. Common ground management: Modal Particles, Illocutionary Negation, and VERUM. In Daniel Gutzmann & Hans-Martin Gartner (eds.), Beyond expressives: *Explorations in use-conditional meaning*, 231–274. Leiden, The Netherlands: Brill.

Roberts, Ian G. 1993. *Verbs and Diachronic syntax: A Comparative History of English and French.* Dordrecht: Kluwer.

Roberts, Ian G., and Anna Roussou. 2003. *Syntactic Change: A Minimalist Approach to Grammaticalization*. Cambridge: Cambridge University Press.

Roberts, Gareth and Betsy Snelle. 2020. Empirical foundations for an integrated study of language evolution. *Language Dynamics and Change* 10, 188–229. DOI: 10.1163/22105832-bja10001

Roberts, Gareth and Maryia Fedzechkina. 2018. Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition* 171, 194–201. DOI: 10.1016/j.cognition.2017.11.005

Shepherd, Susan. 1982. From deontic to epistemic: An analysis of modals in the history of English, creoles, and language acquisition. *Papers from the Fifth International Conference on Historical Linguistics*, ed. Anders Ahlqvist, 316-23. John Benjamins.

Sweetser, Eve. 1990. *From Etymology to Pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge University Press.

Traugott, Elizabeth C. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65: 31-55.

Warner, Anthony R. 1993. *English Auxiliaries: Structure and History: Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press.

Zehr, J. and F. Schwarz (2018). Penncontroller for internet based experiments (ibex).

Zhang, Muye, Maria Mercedes Piñango & Ashwini Deo. 2018. Real-time roots of meaning change: Electrophysiology reveals the contextual-modulation processing basis of synchronic variation in the location possession domain. *40th Annual Conference of the Cognitive Science Society*, 2783–2788.