Simulating semantic change: a methodological note

Our **objective** is to initially illustrate that experimental semantics can be used as a bridge to cross the gap between semantic fieldwork and diachrony as they currently stand. We propose

(1) ***The human diachronic simulation paradigm*** **(HUDSPA)**: Humans confronted with new meaning-form pairings modeled after an attested semantic change will react similarly when they are placed in conditions that resemble those of the actual change (e.g. via a cognate that is similar but did not undergo the transformation investigated).

Background: Diachronic and fieldwork semantics model natural language variation. However, they are often viewed as not (yet) compatible. For instance, Deal (2020) considers variable-force modals and questions diachronic conclusions (e.g. when variable-force is suggested based on 72 Old Engl. examples). We don't engage with OE modals here (other analyses exist, based on Krochian competition, Author xx). But the more general point holds. How can a semantic path of change (especially an unusual one) be established given the impossibility of eliciting contextualized judgments, of receiving comments from interviewed speakers, etc.? **Towards a solution**: One way-out are changes in progress (e.g. D'Arcy 2007); we will not discuss them. There are not enough corresponding changes in progress for many interesting meanings that have arisen historically which have been completed longer ago than, say, two generations. Instead, and even if clear differences exist, we draw from other cases of semantic development in which the extraction of speaker intuitions is disadvantaged, viz. the earliest stages of acquisition (Gleitman et al. 2005). We report in this abstract on **two experiments**: on the development of E(nglish) *even,* simulated from the perspective of G(erman) *eben*, and on the G discourse particle *doch,* through the prism of E *though*. We paid attention to syntax, e.g. by using final *though* in view of relevant factors (van Kemenade 2018). But crucially, in actual history, *eben* didn't develop a meaning such as (Modern) E *even* (G only uses non-cognates of *eben* for additives of improbability), just like *though* did not develop a presuppositional meaning as *doch* (Grosz 2010), regardless of syntax. We have hence used two cues to activate speakers to such readings: one is context to clarify the intended meaning; the other is the instruction to treat the examples as spoken by some non-mainstream G (and E) community and to grade the naturalness of the examples encountered w.r.t. to the context given (on different types of scales; cf. below for a selection of the experiments and analyses conducted). [From an earlier study, we had confirmation that speakers can reliably assign meanings in rich contexts to sentences which they find otherwise inacceptable.] NB: the two experiments are not mirror images of one another. We describe the first one in more detail and the second one primarily w.r.t. the main diverging points below.

Eben **manipulated as** ***even****:* Experimental design: A questionnaire with 12 target items and 13 filler items was used. The target items consisted of 3 item sets with each set consisting of 4 items and respectively licensing readings of *sogar* ('even'), *nur* ('only'), and *auch* ('too/also'). In place of *sogar*, *nur*, and *auch*, the items featured *eben*. All items consisted of a context description and a target sentence as well as a comment section. Subjects were asked to rate the target sentences based on a 7-point scale ranging from 'fully acceptable in context' to 'not at all acceptable in the context'. For the comment section subjects were encouraged to suggest improvements should they find certain expressions odd. Data processing: In total, we had data from 71 subjects, yielding 810 observations (excluding 42 missing values from the ratings). We excluded non-native German speakers. Additionally, we manually categorized the comments provided by the subjects as to their suggestions for improving the target sentences. Based on this manual categorization, we had 199 observations (53 for the *sogar* condition, 94 for *nur*, 52 for *auch*). The criterion here was that the subjects suggested replacing *eben* with *sogar*/*nur*/*auch*. If subjects suggested supplementing *eben* with *sogar*/*nur*/*auch*, commented on an unrelated issue (or did not provide a comment at all) their rating was not considered for this analysis. In descriptive terms, the three conditions were rated as follows:

	sogar 'even'	*nur* 'only'	*auch* ,also/too'
Mean (& median; SD):	5.17 (6; 1.46)	4.34 (5; 1.7)	4.62 (5; 1.83)

For statistical analysis, we transformed the ratings into norm scores and fit the data into a random slope model with *NormScore* as a function of *condition* (3 levels: *sogar*, *nur*, *auch*) allowing for different slopes per *subject*: NormScore ~ condition + (1 + condition | subject).

The estimate for the *sogar*('even')-level is 0.222 and the slope for the *nur*('only')-level is -0.561, for *auch*('also/too') -0.382. In order to obtain a p-value, we conducted a likelihood ratio test, pitching the full model against a null model (i.e. without the factor of interest, *condition*). The three levels of the factor condition affected the transformed ratings (χ^2 (2) = 13.221, p = 0.001346) lowering them by 0.561 for the *nur*-level and by 0.382 for the *auch*-level.

The second experiment, **manipulating *though* as *doch***, was similar in certain respects: 12 target items (joined by 14 fillers) with 4 target items per condition, where the respective readings approximated three different types of particles: *doch, ja, wohl* (cf. Zimmermann 2011 for an overview of the untranslatable material). Some of the main points where this experiment differed: (i) trying to reproduce felicitous readings of particles required the use of slightly longer and only dialogic contexts; (ii) given that the particles do not have counterparts in E, the experiment included two tasks, the first one consisting of a training section and asking if the meaning from the contextual clues was understood; (iii) given that the language in which this experiment was conducted (E) lacks the particles, it could not be expected to have the same precision in the additional comments, but participants were asked to give comments nonetheless. 40 E native speakers have been administered this experiment, but due to the inclusion of attention-testing fillers, only 36 have been considered. The answer to the first task was given through a slider ranging from 1 ('very hard to understand') to 101 ('very easy to understand'); the answer to the second task was a forced-choice *yes/no* slider to test if the item was actually understood. The descriptive statistics for the two tasks of this experiment:

Task 1	*doch*	*ja*	*wohl*
Mean (& SD):	95.40 (10.48)	92.47 (18.53)	92.15 (17.53)
Task 2	*doch*	*ja*	*wohl*
Mean (& SD):	84.45 (25.33)	59.94 (38.34)	23.73 (34.07)

While the sentences seemed easy to understand (Task 1), not all intended meanings were captured reliably. *Doch* readings were rated significantly higher in Task 2 than *ja* and *wohl*.

Discussion: the experiments show that the meanings of the cognates were interpreted more appropriately than competitors. The discourse particle seems calculable from the relationship between the currently available concessive component (reflected in the comments), which is close in meaning to the presupposition of contrast in *doch.* The additive case may seem more surprising. However, if we consider that German *eben* can have e.g. a meaning similar to what Traugott (2006) identifies as a particularizing focus modifier reading (for Early English *even*), as in (2), then we can explain the significantly higher acceptability ratings for the items where *eben* was manipulated for *even*. Traugott describes such a reading of *even* as precursor (Stage II of a 3-stage development) towards the modern one. Given the possible availability of *eben* as in (2) in the subjects' grammars, they might have had an easier time accommodating a scale of (im)probability rather than for items where *eben* was manipulated for *also/too* and *only*. The *only* meaning can be associated with the use of *eben* only in varieties of Austrian G, cf. (3) (subjects evaluated in the experiment reported here: all native speakers of Federal G).

(2) *Peter hat letzte Woche im Krankenhaus Maria kennengelernt. Eben diese Maria hat er heute zufällig im Supermarkt getroffen*.
'Last week, Peter met Maria at the hospital. Today, he ran into exactly that Maria by chance at the grocery store.'

(3) *Ich habe so viel zu tun, aber ich habe eben zwei Hände*.
I have so much to do, but I have only two hands. (Only varieties of Austrian G.)

While points of caution and discussion are in order, HUDSPA at this point shows convergence towards the actually developed meaning if the speakers' grammars are properly factored out. This is a minimal but crucial result towards more refined investigations of change. While, for instance, popular game-theoretic approaches have a similar goal of simulating paths of change, they do so by stipulating (often rather abstract) costs and benefits, so that any course can be attained. HUDSPA, by contrast, constrains the course of change appropriately, by using natural-language intuitions, which can further be probed into experimentally and theoretically.

Selected references: Deal, A.-R. 2020. Comments on diachronic formal semantics (as compared to formal semantic fieldwork). Paper at *Formal Approaches to Grammaticalization*. LSA, N.O., La. || Gleitman, L. et al 2005. Hard words. *Lang Learn & Development* 1: 23-64.