Baseline learners

Rácz, Péter

2023-04-06

Background

A lot of early work in language modelling focussed on regular/irregular variation in the English past tense. This is because the English past tense exhibits both rule-like behaviour (there is a regular rule that applies to the vast majority of form) and analogical behaviour (irregular forms that look similar also behave in a similar way) and because people working on the problem all spoke English.

The English past tense is an unusual example of variation in inflectional morphology, for three reasons. The irregular patterns have very little productivity. There is a wide array of disparate irregular patterns that result from diachronic accidents and do not interact with each other or other parts of the grammar. (Variation in the past tense of irregular forms is social or stylistic but not contextual.) Language user awareness of the pattern is low and usually restricted to specific forms.

Here I look at three inflectional patterns in Hungarian, a language with complex morphology. These patterns exhibit different properties (from the English past tense and one another). I used the Hungarian Webcorpus to build training sets that exhibit variable behaviour. I built test sets based on a forced-choice elicitation experiment using nonce forms. I then went on to implement a range of learning models using the training sets to see how well they predict human responses in the test set.

The learning models were GCM, MGL, fasttext, huBERT (Google's Bert fit on Hungarian Wikipedia and web crawl) (https://huggingface.co/SZTAKI-HLT/hubert-base-cc), and GPT-3 (EleutherAi's GPT-Neox-3 fit on Hungarian Wikipedia and web crawl) (https://huggingface.co/NYTK/PULI-GPT-3SX).

The inflectional patterns

Two patterns relate to verbal conjugation, one to nominal declension. One can apply to all verbs in its category, two have phonological restrictions. One is a social marker, the other two are largely under the radar. One is expressed for a specific function only, the other two across a range of functions.

1. lakok/lakom

The definite form is used in the indefinite in certain 1sg.present verbs in educated colloquial Hungarian. These verbs all end in `-ik` which is originally a reflexive suffix ("fésül" comb.3sg.indef, "fésülök" comb.1sg.indef, "fésülködik" comb.refl.3sg.indef, "fésülködök/fésülködöm" comb.refl.1sg.indef).

More prevalent with some noun-to-verb derivational suffixes (-odik, -zik, -lik). See Rácz 2019. Note that Hungarian verbs is a closed class.

This variation only applies in the 1sg.indef. It has no phonological restrictions. It can apply to all `-ik` verbs. It applies productively to neologisms formed using derivational suffixes that end in `-ik` (e.g. "gugli.zik" google.3sg.indef, "guglizok/guglizom"). The `-m` variant is a marker of educated, careful speech.

2. hotelban/hotelben

Nominal post-positions usually agree with stem in front/back vowel and vowel roundedness. There are front and back stem vowels. Some stem vowels (i,í,e,é) are neutral. (e) and increasingly (é) are front vowels. As a result, back vowel + (e,é) stems can vary. See https://doi.org/10.3765/amp.v8i0.4750 (Gloss: "hotelban" hotel.in)

This variation applies in a wide range of nominal declensions – anywhere where the suffix shows front-back variation. Its rate of use is also sensitive to the suffix. It applies productively to borrowings ("spandexnak" spandex.dat).

3. cselekszik/cselekedik

A Hungarian verb stem can end in CC or CVC. Verbal suffixes can be C- or V-initial. Some stems vary in that a CVC stem is realised with a C-initial suffix and a CC stem is realised with a V-initial suffix. This is likely because we have a CVCV sequence in the former case and a VCCV sequence in the latter case, both of which are phonotactically well-formed (Some verbs always have CVC stems even when this is not phonotactically necessary, other verbs are defective, only have CC stems and some C-initial suffixed forms do not exist.)

Some epenthetic stems are monomorphemic (fürdik, ugrik: ugrani - ugornak - ugranak) but most are formed with a small set of productive derivational suffixes. (Gloss: "cselekszik" act.3sg.indef, "ugrik" jump.3sg.indef)

This variation applies in a wide range of verbal inflections – anywhere where the suffix is consonant-initial. Its rate of use is sensitive to the suffix. It garners very little sociolinguistic attention, apart from some lexicalised forms, such as "emlékszik" remember.3sg.indef / "emlékezik" commemorate.3sg.indef.

Creating the training sets using the Hungarian Webcorpus

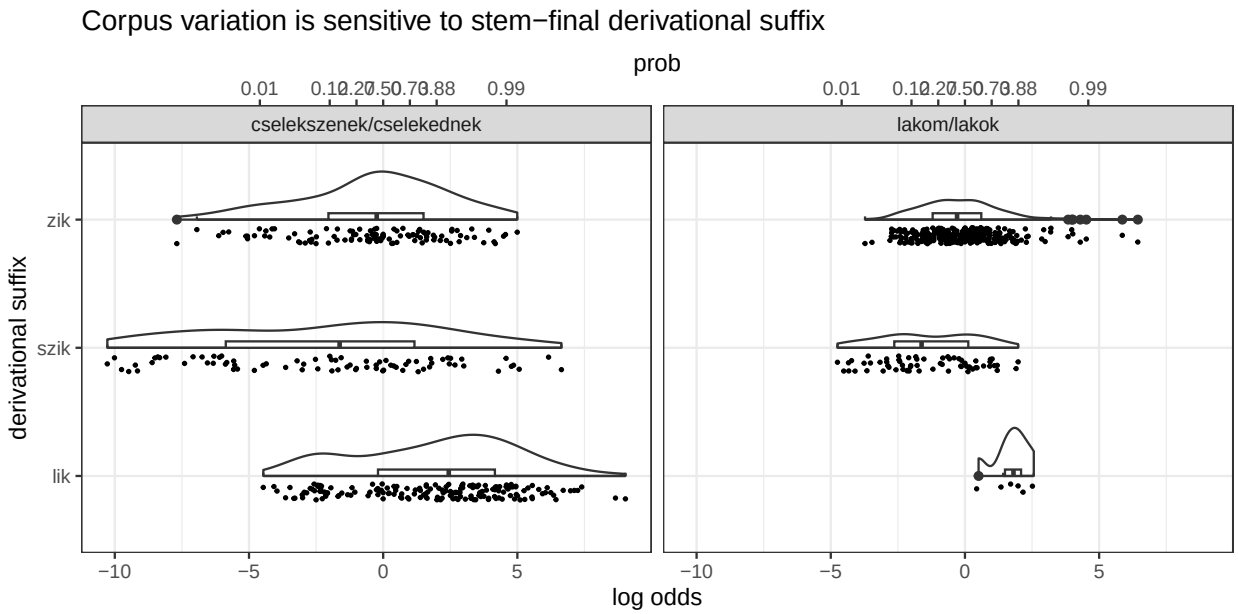
I used the morphologically disambiguated and pos tagged Hungarian Webcorpus 2 (https://hlt.bme.hu/en/resources/webcorpus2) to compile a form frequency list. I used this list to calculate the log odds of variation in the (i) definite/indefinite suffix used in 1sg.indef -ik verb forms, (ii) the front/back suffix used with variable noun stems, and (iii) CVC/CC forms of variable verb+suffix combinations.

For the lakok/lakom 1sg.indef -ik verb forms, there is, in practice, only one variable morphological exponent (whether the 1sg.indef form is realised with the indefinite or the definite suffix). For the hotelban/hotelben variable noun stems, I selected five locative nominal post-positions known to vary: the inessive, the illative, the adessive, the dative, and the sublative. For the cselekszik/cselekedik variable verb stems, I selected five verbal suffixes known to vary: the infinitive, the Pres.NDef.3Pl, the Past.NDef.3Pl, the Cond.NDef.3Sg, and the Pres.NDef.2Pl.

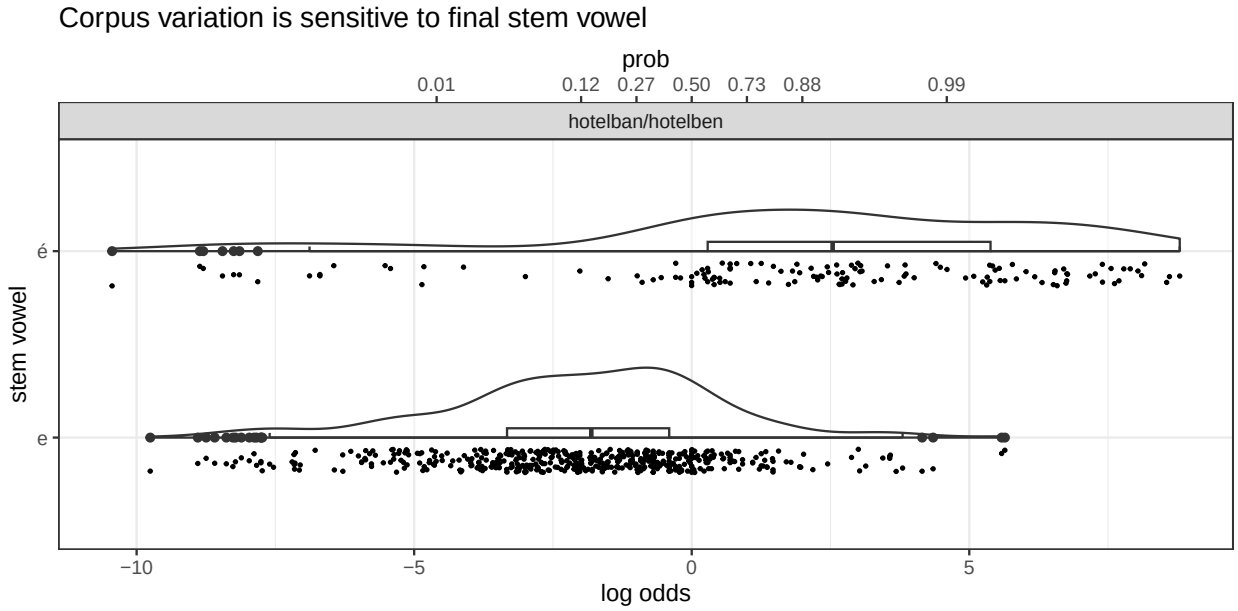
For the nouns, variation does not affect the stem itself, and so identifying variable stems with the five post-positions was straightforward. For the verbs, instead of searching for a list of possibly variable lemmata, each possible CC and CVC variant was built and then grepped individually to make sure the scoop was as large as possible.

All three variable patterns are strongly correlated with cross-referenced sets from the first Hungarian Webcorpus (http://mokk.bme.hu/en/resources/webcorpus/).

The three patterns show structured variation in the resulting dataset. To give one example, the two verbal patterns are sensitive to the type of derivational suffix on the verb stem:



The noun pattern is sensitive to the stem-final vowel:



Creating the test sets

Generating nonce words

An ngram model was built to generate nonce words from pre-defined constituent parts in relevant real word patterns. Nouns were generated based on existing bisyllabic variable noun stems (where a back vowel is followed by e/é). Verbs were generated based on existing bi- and trisyllabic verb stems ending in one of four productive derivational verbal suffixes: -lik, -odik, -szik, -zik. The verb class is closed in Hungarian and new verbs are formed using derivational suffixes: Google - Guglizik, Facebook - Facebookol, etc. Native speakers are far more likely to accept nonce verbs that have recognisable derivational endings.

Existing words were broken up into onset-rhyme couplets and then these were freely recombined to build nonce forms. For verbs, I added some complex onsets to enlarge the set of possible combinations. The recombined set was filtered to make sure it (a) excludes uncommon constituents (e.g. "yiddish" is a Hungarian word but most Hungarian words do not have a "y-" onset) and (b) observes restrictions on syllables and consonant sequences in monomorphemic forms. Most of these restrictions were only relevant for the nouns. If a participant takes the phonotactic cues to parse a nonce noun as a compound, they will be very likely to only consider the second constituent in selecting a suffix (so that all suffixes will be front-only). For verbs, the derivational suffixes constitute a morphemic boundary.

Nonce forms were sampled across relevant dimensions (nouns: is the second vowel e/é; -ik verbs: number of syllables, epenthetic verbs: number of syllables and type of derivational suffix) and then the sampled lists were filtered to make sure (a) they had sufficient edit distance from existing forms and (b) did not start or end with a string identical to an existing word and (c) had sufficient edit distance from one another. Final samples were hand-filtered.

Generating final forms

The final sets consisted of 162 forms per variation type (noun / -ik verb / epenethetic verb). All nouns in the hotelban/hotelben set were bisyllabic, consisting of a back vowel and front e/é. All verbs in the lakok/lakom set were mono- or bisyllabic and ended in -lik (see 'csuklik'), -szik (see 'emlékszik'), or -zik (see 'éhezik'). All verbs in the cselekszik/cselekedik set were mono- or bisyllabic and had one of three alterations: -lik / -ozik (see porlik/porozik, common for loans, see squashol / squashozik), -szik / -dik (see cselekszik/cselekedik), or -zik / -zik (see habzik / habozik).

For the hotelban/hotelben and cselekszik/cselekedik patterns, a suffix was chosen at random for each form. For the lakok/lakom pattern, variation was restricted to one suffix, so this filtering was not necessary.

The baseline experiment

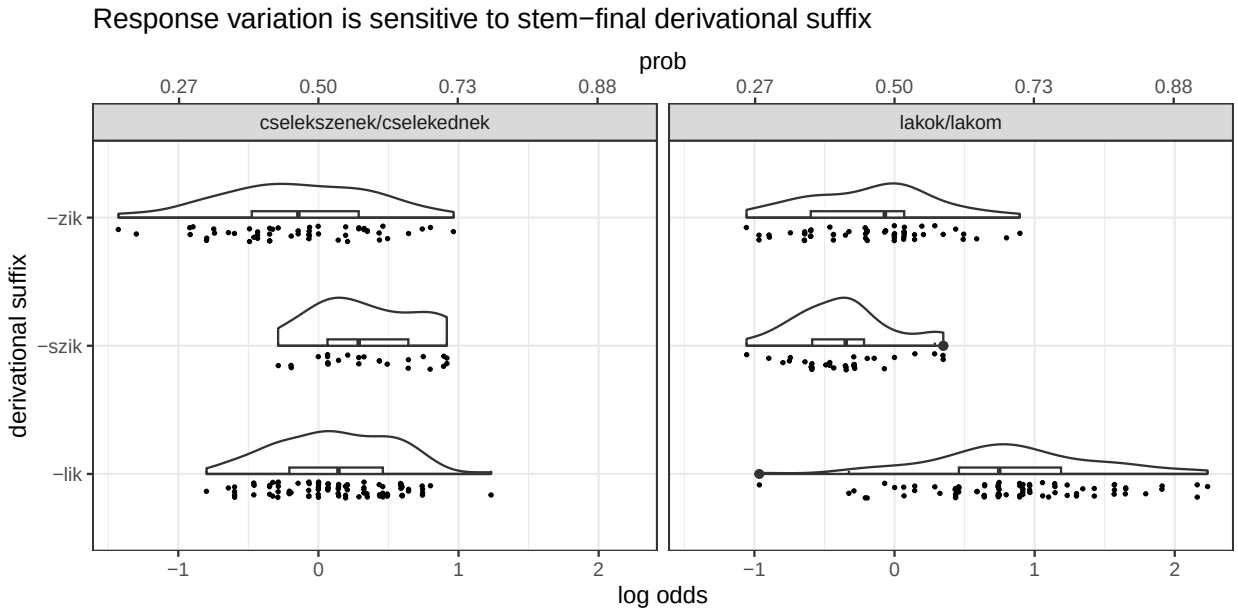
The prompt / target couplets were inserted into simple carrier sentences, following Berko's WUG paradigm (Berko 1958).

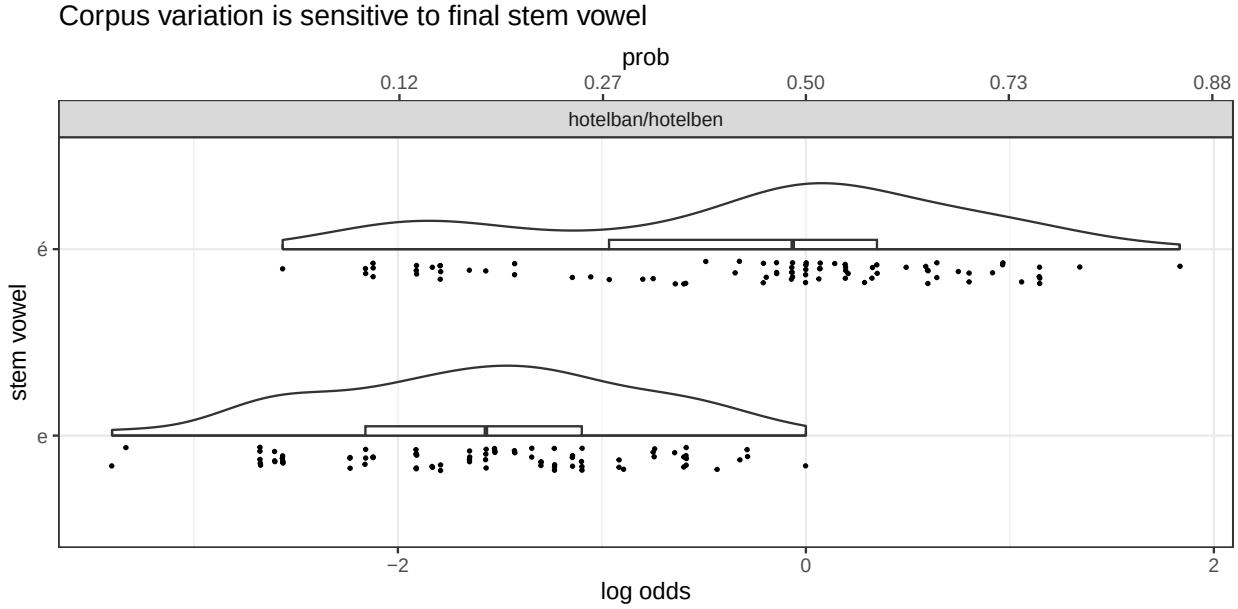
85 students of the Budapest University of Technology and Economics took up the task for course credit in the spring of 2022, 78 finished it (67 women, median age 22). The study was approved by the United Ethical Review Committee for Research in Psychology in Hungary (EPKEB, ref. number 2021-119).

The task was coded in Psychopy (Peirce et al. 2019). Each participant completed the task on their home computer via Pavlovia, an online experimental platform (https://pavlovia.org/). Each participant responded to 162 written prompts with a written binary forced-choice response: 54 for lakok/lakom, 54 for cselekszenek/cselekdnek, 54 for hotelban/hotelben. Participants were instructed that they would see Hungarian words, which might not be familiar to them, and would have to pick a suitable form in the target sentence for the word in the prompt sentence. They were given an example before starting.

Each prompt received a median 27 responses (25th = 26, 75th = 29), for a total of 13284 responses by 78 participants to 486 prompts.

The three sets display structured variation in the test set. This is similar to the variation seen in the training set. This is the plot of log odds across derivational endings in the responses for the verb prompts:



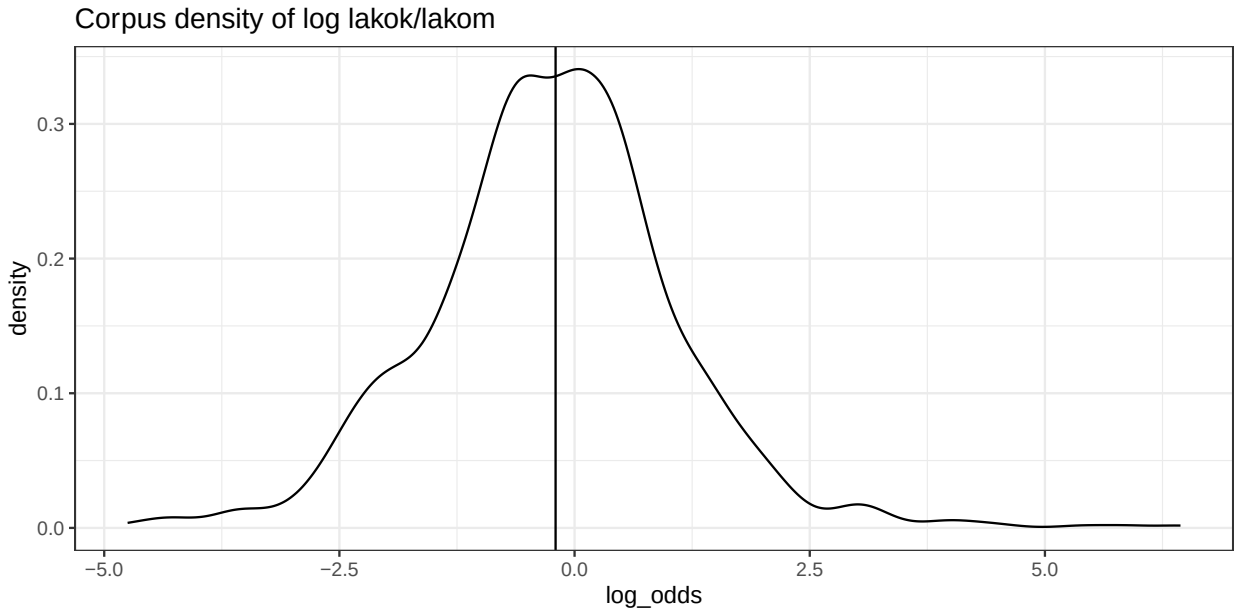


The Generalised Context Model

The generalised context model takes an individual target form and a set of categories. It then calculates the distance of the target form to the training forms in the categories and comes up with a distance to each category. In the present case, the number of categories is always 2 and the distances sum up to 1, where distance to category 1 equals 1 - distance to category 2.

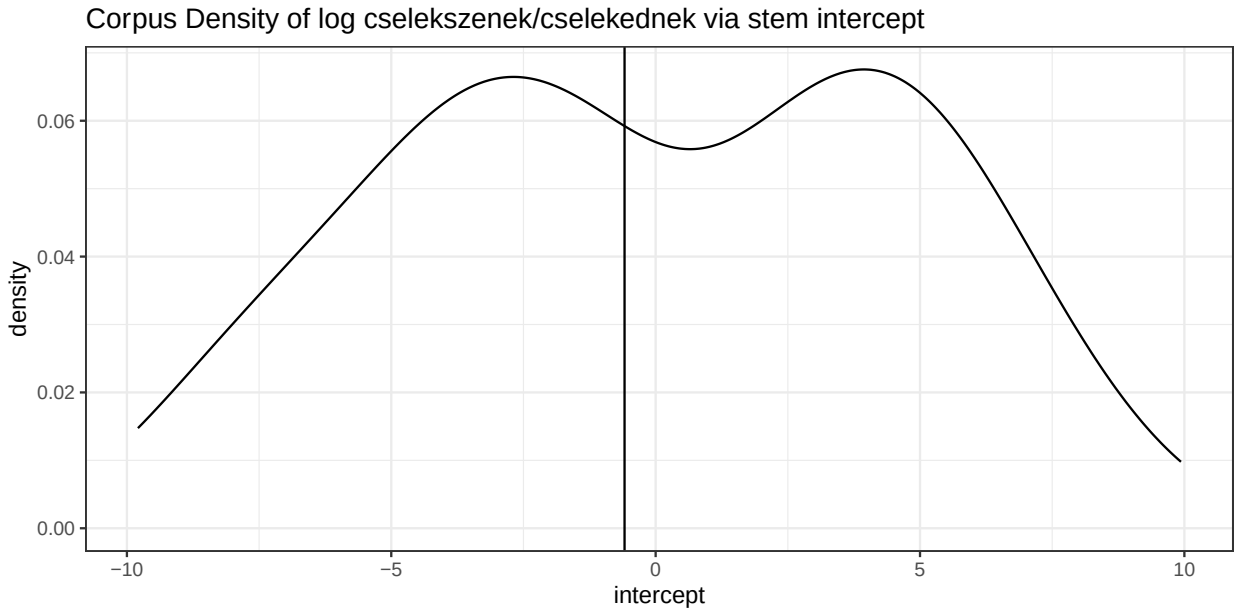
Training sets

For lakok/lakom, all `-ik` verbs are able to show variation. When a verb has no variable forms in the corpus, this is probably because I or somebody else filtered them out at some point or the corpus is too small. But it's better to restrict the training set to attested variable forms.



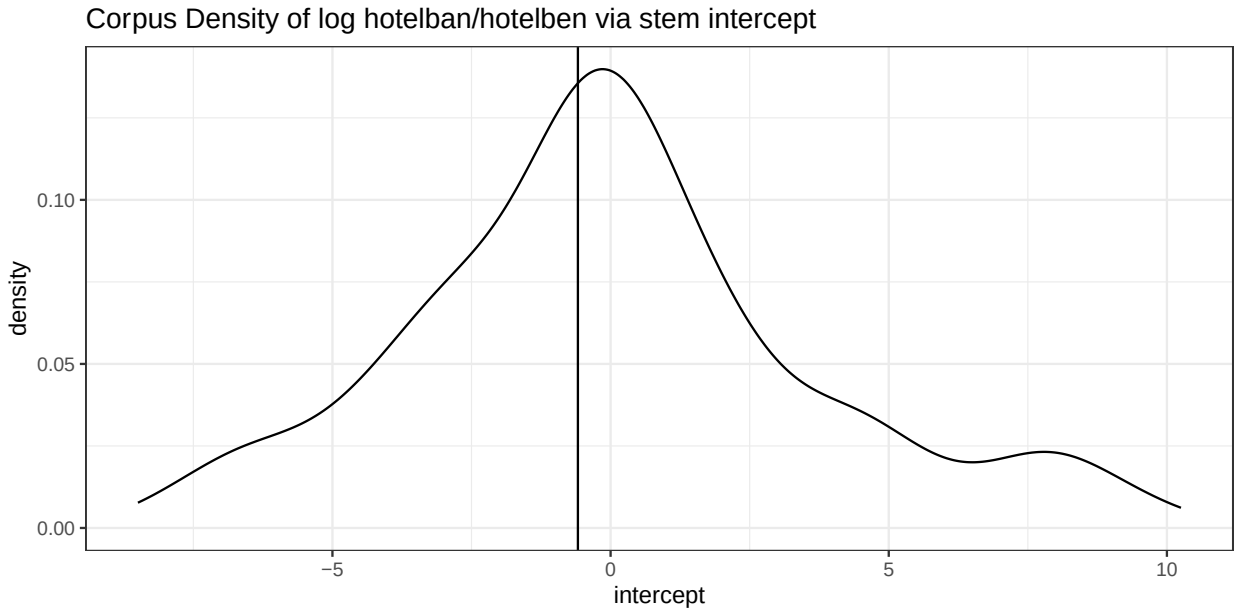
For cselekszenek/cselekednek, there is a set number of stems that are either CVC or CC and some of them vary between the two.

category	n
cc	60
cvc	2046
cvc/cc	111



For hotelban/hotelben, there is a set number of stems that always select front suffixes, a set that always select back suffixes, and a variable set. This is similar to the cselekszenek/cselekednek set. (For the sake of simplicity, we exclude forms that end in a transparent vowel).

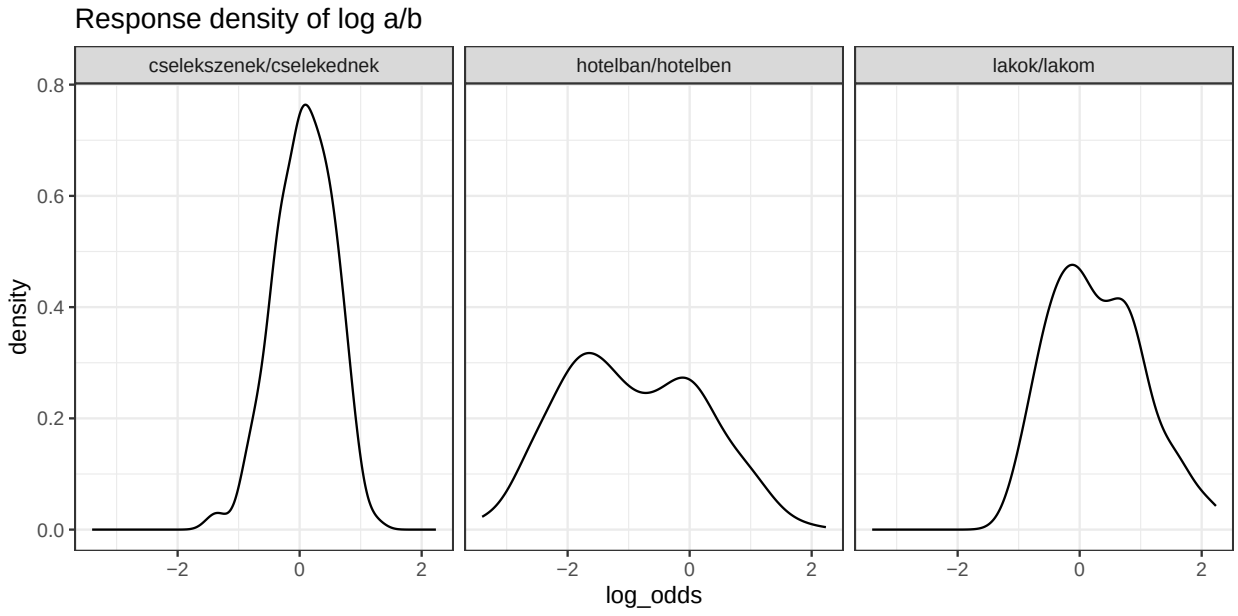
category	n
hotelban	3527
hotelben	3303
hotelban/hotelben	225



For lakok/lakom, I split the variable set at the median. For cselekszenek/cselekednek, I have three sets: the variable set, split at the median, stable CVC and CC forms only, combination of variable and stable forms. For hotelban/hotelben, I have three training sets: the variable set, split at the median, stable front and back forms only, combination of variable and stable forms.

Test sets

The test data are the responses from the experiment. These are aggregated and split down the median as well.



Tuning the model

The GCM has two parameters, s (0-1) and p (0/1). It's worthwhile to search for s.

Minimal Generalisation Learner

I'm using Colin Wilson's CLI for the MGL. (https://github.com/colincwilson/MinimalGeneralizationLearner)

The MGL can think about individual rules as opposed to broad categories. This makes no difference for the lakok group, where there are two possible outcomes, -m and -k. It makes a major difference for the cselekszenek group, where cvc/cc variation is expressed through a range of suffixation patterns. (cselekedik/cselekszik 3sg.indef, cselekednek/cselekszenek 3pl.indef, cselekedünk/cselekszünk 1pl.indef)

I need to build the input files.

```
## # A tibble: 976 x 6
##    base            base_tr         form_1          form_2          freq_1 freq_2
##    <chr>           <chr>           <chr>           <chr>            <dbl>  <dbl>
##  1 ábrándozik      ábrándozik      ábrándozok      ábrándozom         284    610
##  2 adakozik        adakozik        adakozok        adakozom           180    168
##  3 adaptálódik     adaptálódik     adaptálódok     adaptálódom         10      7
##  4 adódik          adódik          adódok          adódom              24      5
##  5 adósodik        adósodik        adósodok        adósodom             2      2
##  6 ágaskodik       ágaskodik       ágaskodok       ágaskodom            8     26
##  7 aggályoskodik   aggá¬oskodik    aggályoskodok   aggályoskodom        4     18
##  8 aggodalmaskodik aggodalmaskodik aggodalmaskodok aggodalmaskodom     60    185
##  9 aggódik         aggódik         aggódok         aggódom           4074  20869
## 10 áhítozik        áhítozik        áhítozok        áhítozom            74    551
## # ... with 966 more rows
```