Stability and change: computational studies of morphology in contact

Micha Elsner, Ohio State University
2023
Thanks to my collaborators

Sara K. Court
(Ph.D. student, OSU)

Andrea D. Sims
(Associate Professor, OSU)

This work is partly supported by NSF:
“Neural discovery of abstract inflectional structure” (Elsner/Sims BCS 2217554)
English in contact

As a major world language, English can’t be understood in isolation. For most speakers, English exists in contact with other languages.
Adult learners of English as a foreign language…

English speakers learning other languages…

And members of multilingual communities.
An old-fashioned view was that the bilingual was “two monolinguals in one person” (Grosjean 1989) whose multiple languages rarely interacted.

There is now strong evidence that multiple languages interact via shared mental representations, in both learners and fluent multilinguals.

Part 1: Lexical analysis to investigate historical contact

Etymology and reanalysis in Maltese

(Court, Elsner, Sims: Intl. Morphology 22; Mediterranean Morphology ‘22)

Part 2: Simulations to understand potential contact outcomes

Celtic nouns under English influence

(in submission)
How does contact affect languages?

One way to tell is to investigate historical outcomes of contact.

Maltese is a Semitic language descended from Arabic, but with heavy influence from Italian and some English.
“Broken” (non-concatenative) plurals

Many Maltese nouns form plurals by changing the way their vowels are arranged.

This process original in Arabic, but applies to words of multiple origins in Maltese.

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
<th>Gloss</th>
<th>Allomorph</th>
</tr>
</thead>
<tbody>
<tr>
<td>fardal</td>
<td>fradal</td>
<td>'apron'</td>
<td>CCVVCVC</td>
</tr>
<tr>
<td>birra</td>
<td>birer</td>
<td>'beer'</td>
<td>(C)CVCVC</td>
</tr>
<tr>
<td>kbir</td>
<td>kbar</td>
<td>'big'</td>
<td>CCVVC</td>
</tr>
<tr>
<td>ftira</td>
<td>ftajjar</td>
<td>'type of bread'</td>
<td>CCVjjVC</td>
</tr>
<tr>
<td>bitha</td>
<td>btieli</td>
<td>'yard'</td>
<td>CCVVCV</td>
</tr>
<tr>
<td>sider</td>
<td>isdra</td>
<td>'chest'</td>
<td>VCCCV</td>
</tr>
<tr>
<td>marid</td>
<td>morda</td>
<td>'sick person'</td>
<td>CVCCV</td>
</tr>
<tr>
<td>ghodda</td>
<td>ghodod</td>
<td>'tool'</td>
<td>(gh)VCVC</td>
</tr>
<tr>
<td>elf</td>
<td>elf</td>
<td>'thousand'</td>
<td>VCCV</td>
</tr>
<tr>
<td>gharef</td>
<td>ghorrief</td>
<td>'wise man'</td>
<td>CVCCVVC(V)</td>
</tr>
<tr>
<td>ghama</td>
<td>ghomja</td>
<td>'blind person'</td>
<td>(gh)VCCV</td>
</tr>
</tbody>
</table>

Table 6: Broken plural allomorphs in Maltese, from Nieder et al. (2021a)
‘Hybrid’ or ‘stratal’ models of the lexicon

Foreign words (extragrammatical)

Partly-nativized words (grammar I)

Native words (grammar II)

foreign affixes or “elsewhere” class

perhaps restricted set of affixes

loosely based on Ziani 2020 for loans in Arabic; Spagnol 2011, Borg and Gatt 2017 in Maltese
Concatenative plurals are also a complex set

Maltese has multiple plural suffixes of different origins.

Deciding which suffix goes with which word is not necessarily simple!
Analogical models of the lexicon

Common words remain stable because learners have opportunities to observe the correct plural.

Unfamiliar words are modeled with reference to existing ones, which compete to provide alternatives.

See Fertig 2017 for an overview of analogical models
Predictions?

How a word sounds is informative about its etymology, and vice versa.

**Both accounts** predict both of these should matter.
Predictions?

How a word sounds is informative about its etymology, and vice versa.

**Both accounts** predict both of these should matter.

**Stratal accounts** suggest:

Easier to divide concatenative from non-concatenative (across strata) than to pick the specific process (within stratum).

Surface form mostly redundant once etymology is known.
Predictions?

How a word sounds is informative about its etymology, and vice versa.

**Both accounts** predict both of these should matter.

**Analogical accounts** suggest:

- No high-level split between concatenative and non-concatenative.
- Surface form highly informative despite knowing etymology.
Methods

We use information theory to quantify the predictive power of our various factors.

Following Williams et al. 2020 “Predicting declension class from form and meaning”
<table>
<thead>
<tr>
<th><strong>Given:</strong></th>
<th><strong>Predict:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface form</td>
<td>Concatenative or non-concatenative?</td>
</tr>
<tr>
<td>Surface form and etymology</td>
<td>Concatenative or non-concatenative?</td>
</tr>
<tr>
<td>Surface form</td>
<td>Specific plural type</td>
</tr>
<tr>
<td>Surface form and etymology</td>
<td>Specific plural type</td>
</tr>
</tbody>
</table>

(*we include gender as a covariate in all analyses*)
total uncertainty about concatenation

.81 bits of total uncertainty
.81 bits of total uncertainty

etymology and surface form redundantly contribute 6% of this amount
.81 bits of total uncertainty

etymology and surface form redundantly contribute 6% of this amount

etymology contributes up to 13%
.81 bits of total uncertainty

etymology and surface form redundantly contribute 6% of this amount

etymology contributes up to 13%

surface form contributes up to 21%
.81 bits of total uncertainty

2.65 bits of total uncertainty
.81 bits of total uncertainty

2.65 bits of total uncertainty

etymology and surface form redundantly contribute 15% of this amount

etymology contributes up to 22%

surface form contributes up to 42%
Conclusions

Both etymology and surface form contribute non-redundantly to Maltese plural prediction.

Surface form is more informative than etymology.

Fine-grained information about the word is more informative about the specific plural type than whether the plural is concatenative:

No evidence for a high-level split!
History is the outcome, but what’s the process?

Maltese came about through extensive multilingual contact, but that contact happened in the past.

We can also observe contact reshaping systems in the present.
Scottish Gaelic language shift (Dorian 1978)

Nancy Dorian studied a community in which Gaelic was being replaced by English.

Multiple plural markers remain even among heritage speakers (the youngest, non-fluent generation), but some types of plural are almost entirely lost.

In particular, suffixes seem to survive better than non-suffixes.
“[These share] the fact that the device in question (quantity change) plays no role whatever in English pluralization” - Dorian 1978

Is Dorian right that we can attribute these effects to English contact? How could we (begin to) tell?
Comparisons

As Jarvis (2000) points out, it would be methodologically ideal to compare across populations for which everything except language background was held constant.

For instance, it would be very convenient to compare English-dominant speakers to speakers whose native language was Arabic or Maltese, for whom non-suffix plurals are common.

But this is rarely possible! Exposure, social attitudes, and learning materials usually all differ.
Can simulations tell us about possible outcomes?

Model learns to predict from training words and language tags

We measure its uncertainty on unseen words

transformer neural network

child
LANG_EN
more stuff

children

model from Elsner and Court 2022; see also Elsner 2020
Compare simulated learners

Due to data shortage, we use Irish rather than Scottish Gaelic.

We compare:

“Monolingual” (Irish data only)

“L1 English” (28k English nouns, limited Irish)

“L1 Maltese” (28k Maltese nouns, limited Irish)
Irish noun plurals

<table>
<thead>
<tr>
<th>Suffixal allomorphs: 58%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ola ~ olaí</td>
<td>‘oil’</td>
</tr>
<tr>
<td>muc ~ muca</td>
<td>‘pig’</td>
</tr>
<tr>
<td>inne ~ inní</td>
<td>‘guts’</td>
</tr>
<tr>
<td>rud ~ rudaí</td>
<td>‘thing’</td>
</tr>
<tr>
<td>góraí ~ góraithe</td>
<td>‘goal’</td>
</tr>
<tr>
<td>idé ~ idéanna</td>
<td>‘idea’</td>
</tr>
<tr>
<td>jib ~ jibeanna</td>
<td>‘jib’</td>
</tr>
<tr>
<td>lao ~ laonna</td>
<td>‘calf’</td>
</tr>
<tr>
<td>others ...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stem allomorphs: 35%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bac ~ baic</td>
<td>‘barrier’</td>
</tr>
<tr>
<td>sáach ~ sáigh</td>
<td>‘well-fed pers’</td>
</tr>
<tr>
<td>fear ~ fir</td>
<td>‘man’</td>
</tr>
<tr>
<td>Síneach ~ Sínigh</td>
<td>‘Chinese pers’</td>
</tr>
<tr>
<td>mac ~ mic</td>
<td>‘son’</td>
</tr>
<tr>
<td>others ...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suffix plus stem: 6%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cúil ~ cúlacha</td>
<td>‘corner’</td>
</tr>
<tr>
<td>others ...</td>
<td></td>
</tr>
</tbody>
</table>
Hypotheses

1. Monolingual models will lag behind with least Irish exposure (boring technical reasons: models perform poorly with little data)
2. L1 English model will have an early advantage with suffixes (following Dorian and other language learning studies)
3. L1 Maltese model will have an early advantage with non-suffixes
4. Models will eventually converge given enough Irish (effects of language transfer are strongest for least-fluent speakers; MacWhinney 2005, among others)
Lang: Plural

- L1 Eng: Stem
- L1 Mlt: Stem
- Mono: Stem
- L1 Eng: Suffix
- L1 Mlt: Suffix
- Mono: Suffix
English models do best on suffix plurals.
Maltese models do best on non-suffix plurals
Results grow more similar with more Irish exposure
Maltese causes interference as well as facilitation for non-suffix plurals.
Some additional results

L1 English and Maltese models don’t just produce more suffixal and non-suffixal plurals (respectively)...

But are better at producing the correct type.

We conjecture it has learned to attend more to cues at word ends or within stems.

The model produces unmarked (zero) plurals, which Dorian also observes.
Conclusions

Computational models can be used as simulated language learners: they undergo *facilitation* and *interference* based on language contact, just like humans.

This makes them useful for hypothesis generation in situations where controlled studies are hard to run.

English L1 influence is consistent with the results Dorian reports; these asymmetries in which patterns are lost are caused by the (suffixal) nature of English, not just by limited exposure to Irish.
We favor analogical models of contact

Evidence against models in which learners maintain multiple grammars or coherent lexical strata for the different languages they speak.

Rather, individual words within the system are flexibly assigned to plural types based on cues shared across the linguistic system.

Thus, effects we find or predict are often strongly lexicalized, and predictable from surface word form.
English does not exist in a vacuum. When we imagine English speakers, we should consider what other languages they know, and how contact influences flow in both directions.

Thank you!
Matching human ratings of “wug” words

2021 SigMorphon shared task