

# Computational morphology

# The early history of computational morphology

# The fundamental theorem of computational morphology

Software which processes language should process words not as atomic units, but as the intricately structured objects linguists recognize them to be.

# Early days, and the AI winter

The linguistic potential of digital computing was recognized almost immediately after ENIAC, the first digital computer, debuted in 1945.

Vast amounts of Cold War defense funding was poured into Russian-to-English machine translation...

...until the 1966 ALPAC report, which concluded that very little progress had been made. It instead recommended, among other projects:

- the development of methods for automatic evaluation of machine translation and comparison with translations generated by human experts, and
- the construction of digital dictionaries.

# Word processing

Much of the work from the late '60s through the 1980s' personal computing boom was focused on word processing technologies, including

- digital dictionaries,
- spell checkers,
- typesetting tools, and
- input method engines (IMEs) for non-alphabetic scripts.

# Digital dictionaries (1/)

Even small dictionaries could overwhelm the limited memory of early PCs, though.

E.g., a list of English headwords from the 1933 Webster's dictionary is far larger than the capacity of a 3.5" 1.44 MB floppy disk (first sold 1986).

And the problem is far more acute if one has to store all inflected forms.

E.g., in English, one has to store at most two forms of a noun (the singular and plural) and up to six forms of a verb.

## Digital dictionaries (2/)

E.g., in English, one has to store at most two forms of a noun (the singular and plural) and up to six forms of a verb.

But in Russian, there are up to twelve inflectional variants of every noun (two numbers times six cases, minus syncretisms) and dozens of forms of every verb.

And in Archi, a language spoken in Dagestan, there are more than 1.5 million possible forms of each verb (Kibrik 1998).

# Lexicon compression

Perhaps unsurprisingly, one highly-effective way to compress a digital lexicon is to generate inflected forms by applying word formation rules to a list of stems.

An early example of this is the MITalk system (Allen 1987), an English text-to-speech engine.\* It "generates" complex word by concatenating (orthographic) stems and affixes. E.g., for *scarcity*; MITalk provides two decompositions:

- *scarce* + *-ity*, with a *e*-deletion rule, or
- *scar* + *city*, which would lead to an incorrect pronunciation.

MITalk could generate the pronunciations (and analyses) of over 100,000 words using just 12,000 stems, affixes, and adjustment rules (Klatt 1987: 773).

\*The commercial version, DECtalk, was the "voice" of the late Stephen Hawking.



# Data sparsity

Data sparsity is also a major issue for morphologically-aware text processing.

In any finite text sample, we are likely to see only a small fraction of the words speakers can generate (Lignos & Yang 2016).

This is a corollary of *productivity* in the Humboldt-Chomsky sense.

Thus, if you search the internet for *fishing*, you might also be interested in documents that mention *fish* or *fishers*, even if they do not mention *fishing* itself.

# Morphological conflation

We therefore may wish to conflate related words when *indexing* (i.e., ingesting) web pages, and also when *searching* (i.e., matching queries to documents).

One heuristic approach is *stemming* (Porter 1980), in which language-specific rules are applied to each word to remove common affixes like *-ing*, *-er* and *-s*.

Unlike the MITalk, stemming does not require a lexicon of stems and affixes, just a cascade of rules.

According to various unconfirmed sources, Google search started using stemming around 2003.

# Ad hoc-ness

Early spelling correction engines (e.g., McIlroy 1982) saved space by permitting fanciful analysis.

E.g., *forest* could be "derived" by affixing the superlative suffix *-est* to *fore*.

E.g., a stemmer doesn't need to produce real words, or even linguistically valid "stems"; it just needs to create semantically coherent equivalence classes (`produc`, `semant`, `equival`).

Many of these early techniques take advantage of English's relatively-impoverished, generally concatenative inflectional system, and it's not clear whether they could be generalized to languages with richer morphologies.

# The modern era

Most of the aforementioned problems, including

- compression/storage
- sparsity
- conflation
- ad hoc, almost-good-enough solutions

persist, despite the fact that computers have many orders of magnitude more computing power than they did in the 1980s.

# Tasks in computational morphology

# Morphological tagging

In the simplest setting, we simply want to obtain a detailed morphological summary of a given word.

E.g., for the word *puppies* this summary include that it's a noun, that it's plural, etc.

This information may not seem terribly useful on its own, but is extremely valuable for "downstream" tasks, such as parsing.

E.g., in German, which has a relatively rich inflectional system and relatively free word order, parsing is extremely difficult without morphological tagging (e.g., Fraser et al. 2013).

# Segmentation

A morphological tagging summary might also include the word's decomposition, its *segmentation*.

E.g., it might be written `puppy+s`, if `+` is used to indicate morpheme boundaries.

This gives us the *lemma* (citation form) and a list of affixes.

# Lemmatization

Or, rather than tagging and segmenting a word, we can replace it with its lemma.

This is, ideally, a more linguistically-informed approach to morphological conflation.



# Unsupervised segmentation

**Task:** given a **word list** (and their **frequencies**), automatically **segment** the words.

Morpho Challenge 2005 (Creutz & Lagus 2005):

6755	sea	sea
1	seabed	sea bed
1	seabeds	sea bed s
2	seabird	sea bird
34	seaboard	sea board
1	seaboards	sea board s

**Results:** F1-scores up to the '70s for English, worse for Finnish, German, and Turkish.

# Analysis in context

**Task:** given a **sentence**, generate the **lemma** and **inflection** for each word.

SIGMORPHON 2019 shared task (McCarthy et al. 2019), sub-task 2:

They	they	N;NOM;PL
buy	buy	V;SG;1;PRS
and	and	CONJ
sell	sell	V;PL;3;PRS
books	book	N;PL
.	.	PUNCT

**Results:** overall lemma and inflection in the high '90s.

# Analysis: in isolation vs. in context

Tagging, segmentation, and lemmatization are together known as *morphological analysis*.

Analysis can be performed in isolation (i.e., one word at a time), or we can resolve morphological ambiguities using the broader linguistic context (i.e., nearby words in a sentence).

E.g., the word *cooler* might receive different analyses in contexts such as *the \_\_ was full of beer and bait ...* vs. *the English daisy prefers \_\_ weather ...*

# Morphological generation

The inverse problem, generating the appropriate morphological form given a morphological specification (or sentential context) is known as *generation*.

This is a key component of digital assistants like Cortana, Siri, and the Google Assistant.

# Natural language generation (1/)

```
weather {  
  temp: 53  
  conditions: CLOUDY  
  location_name: "Brooklyn"  
}
```

Template: It's \$temp degrees and \$conditions in \$location.

But note when \$temp = 1, we use *degree* instead of *degrees*.

This gets a lot harder real fast.

# Natural language generation (2/)

Consider Russian:

один градус

'one degree'

два градуса

'two degrees'

пять градус**ов**

'five degrees'

двадцать три градуса

'twenty-three degrees'




двадцать пять градус**ов**

'twenty-five degrees'

- If the last term in the number name is *singular*, we use the nom.sg.
- If the last term in the number name is *paucal* [2-4], we use the gen.sg.
- If the last term in the number name is *plural* [5-9], we use the gen.pl.

"None pizza with left beef"

### Choose Toppings

Toppings	 WHOLE	 LEFT	 RIGHT	<b>NONE</b>	AMOUNT
Cheese	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Sauce	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Pepperoni	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Extra Large Pepperoni	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Italian Sausage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Green Peppers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Black Olives	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Pineapple	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Mushrooms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Onions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
<b>Beef</b>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Normal ▾
Ham	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Bacon	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	
Anchovies	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	



# Inflection generation

**Task:** given a word in citation form (a *lemma*) and a morphological specification (an *inflection*), generate the appropriate *inflected form*.

CoNLL-SIGMORPHON 2017 shared task (Cotterell et al. 2017), subtask 1,  
CoNLL-SIGMORPHON 2018 shared task (Cotterell et al. 2018), subtask 1:

English:	permaban	V;V.PTCP;PST	permabanned
German:	Stiefvater	N;ACC;PL	Stiefväter
Hungarian:	homeosztázis	N;DAT;PL	homeosztázisoknak
Russian:	сарафанное радио	N;ESS;PL	сарафанных радио



# Paradigm cell filling

**Task:** given a *lemma* and an incomplete inflectional *paradigm*, generate the **remaining cells of the paradigm**.

CoNLL-SIGMORPHON 2017 shared task (Cotterell et al. 2017), sub-task 2:

release	release	V;NFIN
release	releases	V;3;SG;PRS
release	releasing	V;V.PTCP;PRS
release	released	V;PST
release	released	V;V.PTCP;PST

**Results:** overall per-form accuracies in the '80s.

# Paradigm generation

**Task:** given a *lemma*, generate the **full inflectional paradigm**.

Dreyer & Eisner 2011, Durrett & DeNero 2013:

schleichen	schleiche	V;1;SG;PRS
schleichen	schleichst	V;2;SG;PRS
schleichen	schleicht	V;3;SG;PRS
schleichen	geschlichen	V;V.PTCP;PST

**Results:** overall per-form accuracies in the '90s (but only 3 languages, and predates modern neural network modeling...)

# Reinflection

**Task:** given an **inflected form** (**not** a lemma) and an **inflection** generate the desired **inflectional form**.

CoNLL-SIGMORPHON 2017 shared task (Cotterell et al. 2017), sub-task 2:

gekommen

V;2;PL;PRS

kommt

Milchprodukt

N;NOM;PL

Milchprodukte

**Results:** overall per-form accuracies in the '90s.

# Generation in context

**Task:** generate the correct **inflection form** of a **lemma** using just **sentential context**.

SIGMORPHON 2018 shared task (Cotterell et al. 2018), sub-task 2:

The dogs (dog) are barking .

**Results:** overall low per-form accuracies ('40s-'50s).

# Computational morphological modeling

# Three cultures

1. Knowledge-based finite-state analyzer-generators (see Gorman & Sproat ch. 6, assigned reading).
2. Data-driven neural network taggers and sequence-to-sequence models.
3. Hybrids of the two.

# A defense of orthographic representation (1/)

We generally assume that inputs and outputs for analysis and generation are represented in orthographic rather than phonemic or phonetic form, largely as a matter of convenience.

For some languages, for instance Spanish or German, this decision has little impact, because the language uses a “shallow”, highly-consistent orthography which is quite close to phonemic representation.

## A defense of orthographic representation (2/)

On the other hand, English or Korean, use a “deep” orthography in which the relationship between spelling and pronunciation is more abstract. This reduces the need to model any pronunciation variation not indicated in written form.

E.g., For example, English spelling does not generally indicate changes in vowel quality triggered by the addition of derivational suffixes like *-ity*. Thus *sane* /sejn/ and *sanity* /sæ.nɪ.ti/, for instance, are spelled more similarly than they are pronounced (Chomsky and Halle 1968: 44f.). Thus we do not need to model this vowel change in written English.



Weird inflects but



# Weird inflects but OK: making sense of morphological generation errors

Kyle Gorman<sup>\*</sup>, Arya D. McCarthy<sup>†</sup>, R. Cotterell<sup>†</sup>, Ekaterina Vylomova<sup>‡</sup>, Miikka Silfverberg<sup>§</sup>, and Magdalena Markowska<sup>\*</sup>

\*: Graduate Center, City University of New York

†: Johns Hopkins University

‡: University of Melbourne

§: University of Helsinki

# Inflection generation

**Task:** given a word in citation form (a *lemma*) and a morphological specification (an *inflection*), generate the appropriate *inflected form*.

CoNLL-SIGMORPHON 2017 shared task (Cotterell et al. 2017), subtask 1,  
CoNLL-SIGMORPHON 2018 shared task (Cotterell et al. 2018), subtask 1:

English:	permaban	V;V.PTCP;PST	permabanned
German:	Stiefvater	N;ACC;PL	Stiefväter
Hungarian:	homeosztázis	N;DAT;PL	homeosztázisoknak
Russian:	сарафанное радио	N;ESS;PL	сарафанных радио

# Our contribution

1. An error taxonomy for inflectional generation errors
2. A manual error analysis for the CoNLL-SIGMORPHON 2017 shared task

# The systems

**UE-LMU-1** (Bergmanis et al. 2017): recurrent neural network (RNN) with a bidirectional gated recurrent unit (GRU) encoder, a unidirectional GRU decoder, and a standard *soft attention mechanism* (cf. Kann and Schütze 2016); also uses some data augmentation tricks.

It is ranked the best overall (macro avg.: 95.23%).

**CLUZH-7** (Makarov et al. 2017): a neural encoder-decoder using a *hard monotonic attention mechanism* with special edit operations (Aharoni and Goldberg 2017).

It is ranked the second best overall (macro avg.: 95.12%) and best on eight languages including Hungarian and Spanish.

# Data sources

Data in the shared task is sampled from UniMorph (Kirov et al. 2016, 2018), a free morphological database. That data in turn is largely extracted from Wiktionary (<https://www.wiktionary.org/>), a free collaborative dictionary.

In the shared task, data is sampled using wordform frequencies from Wikipedia, weakly approximating the statistical properties of the primary linguistic data.

Systems were evaluated in low (100 triples), medium (1,000 triples), and high (10,000 triples) data conditions. We focus on the high data condition.

# Developing the taxonomy

Since at least Rumelhart and McClelland (1986), there has been interest in whether inflection generation errors can be given linguistic characterizations (see, e.g., Pinker and Prince 1988, Sproat 1992:216f.); e.g., \**membled* for *mailed*.

Kirov and Cotterell (2018) claim that modern neural network architectures generalize well while largely eliminating these bizarre errors; Corkery et al. (2019) argue that their model's predictions still align poorly with human productions.

# The error taxonomy

- TARGET:
  - FV (free variation): multiple wordforms are permitted, but only one is present in the gold data
  - UNIMORPH (extraction): errors in the UniMorph extraction procedure
  - WIKTIONARY: errors in the forms listed on Wiktionary
- SILLY: "bizarre" errors that defy linguistic categorization
- ALLOMORPHY: misapplication of existing (i.e., independently attested) allomorphic patterns in the target language
- SPELLING: misapplication of language-specific spelling rules

(NB: we also assume TARGET errors are non-specific to the model.)



# Inter-annotator agreement

<b>Language</b>	<b>RA</b>	<b>Krippendorff's <math>\alpha</math></b>
Dutch	.949	.907
English	.861	.855
Spanish	.861	.875

Language	Noun	Verb	Adjective	UE-LMU-1	CLUZH-7	Overlap
Dutch	x	✓	✓	31	32	84%
English	x	✓	x	28	32	24%
Finnish	✓	✓	✓	49	65	44%
German	✓	✓	x	70	88	48%
Hungarian	✓	✓	x	136	132	65%
Italian	x	✓	x	21	24	50%
Latin	✓	✓	✓	187	190	56%
Polish	✓	✓	✓	72	79	74%
Portuguese	x	✓	x	9	10	73%
Romanian	✓	✓	✓	109	122	59%
Russian	✓	✓	✓	84	79	60%
Spanish	x	✓	x	27	25	44%

# Target errors

These are very common in certain languages:

- Free variation errors:
  - Finnish: *omenoiden*, *omenoitte*, *omenojen*, *omenien*, *omenain* are all possible gen.pl. forms of *omena* 'apple' but only one is found in UniMorph
- Extraction errors:
  - Hungarian: many triples with the wrong case features
  - Latin: all lemmata missing macrons
  - Romanian: a header reading "definite articulation" incorrectly taken as an inflected form
- Wiktionary errors:
  - Spanish: *\*demarce* (cf. *demarque*) is given as the 1sg. pres. subj. of *demarcar* 'to demarcate'

# Silly errors

These were relatively rare overall, though somewhat more common for UE-LMU-1 than for CLUZH-7:

- German (CLUZH-7): \**Schädling**s**bekämpfung**s**mit* for the gen.pl. of *Schädling**s**bekämpfung**s**mittel* 'pesticide'
- Latin (UE-LMU-1): \**praes**ō**s* for the acc.pl. of *praes**ul*** (a title used by various Roman religious and political leaders)
- Russian (UE-LMU-1): \**prinuditel'**n**ym abótam* for the dat.sg. of *prinuditel'**n**ye rabóty* 'forced labor'
- Spanish (UE-LMU-1): \**at**u**engáis* for the 2pl. pres. subj. of *atener* 'to maintain'

# Allomorphy errors

- Stem-final vowels in Finnish
- Ablaut in Dutch and German
- Umlaut in German
- Consonant gradation in Finnish
- Linking vowels in Hungarian
- Yers in Polish
- Spanish diphthongization
- Noun plural suffixes in German
- Genitive singular suffixes in Polish
- Verbal prefixes in German
- Animacy in Polish and Russian
- Aspect in Russian
- Vowel harmony in Finnish compounds
- Internal inflection in Russian compounds

# Allomorphy errors

- Stem-final vowels in Finnish
- **Ablaut in Dutch and German**
- Umlaut in German
- Consonant gradation in Finnish
- **Linking vowels in Hungarian**
- **Yers in Polish**
- Spanish diphthongization
- Noun plural suffixes in German
- **Genitive singular suffixes in Polish**
- **Verbal prefixes in German**
- Animacy in Polish and Russian
- **Aspect in Russian**
- **Vowel harmony in Finnish compounds**
- Internal inflection in Russian compounds

# Ablaut in Dutch and German

Overapplication (e.g.):

Dutch \**pront* for *printte* 'printed' (ppl.)

German: \**versächten* for *versenkten* 'he would have sank'

Underapplication (e.g.):

German: \**sauft* for *soff* 'I/he/she/it was drinking'





# Linking vowels in Hungarian

The Hungarian noun plural suffix *-k* is usually preceded by a linking vowel, one of *-a-*, *-o-*, *-e-*, or *-ö-*; e.g., *vérek* from *vér* 'blood'.

Whether a [+back]-controlled stem selects *-a-* or *-o-* is largely unpredictable (Siptár and Törkenczy 2000:224f., Vago 1980:110f.). Thus we see many errors (e.g.):

\**masszázsakból* for *masszázsokból* (relative pl. of *masszázs* 'massage')

# Yers in Polish

Yers refer to the "fleeting vowels" found in many Slavic languages.

Both position and quality (i.e., backness) of the fleeting vowels are unpredictable: they cannot be analyzed as epenthetic. They are therefore usually assumed to be underlyingly present but distinct from other vowels (Lightner 1965, Gussman 1980:36f., Rubach 1984:41, 1986), with "lowering" (i.e., surface realization) conditioned by other yers. Naturally this contributes to many errors (e.g.):

\**klęsek* for *klęsk* gen.pl. (from *klęska* 'defeat')

\**żagieli* for *żagli* gen.pl. (from *żagiel* 'sail')

# Genitive plural suffixes in Polish

Polish has two gen.sg. suffixes *-a* and *-u*, and it is generally impossible to predict which a given stem will select (Dąbrowska 2001, 2005, Kottum 1981, Maunsch 2003). This leads to many errors (e.g.):

\**ateuszu* for *ateusza* 'atheist'

\**legaru* for *legara* 'joist'

\**krzyka* for *krzyku* 'scream'

\**izotopa* for *izotopu* 'isotope'

# Verbal prefixes in German

- The *separable* prefixes separate from their host when tensed
- The *inseparable* prefixes are always prefixed to their host verb
- Some prefixes, such as *um-*, can be separable or inseparable, depending on the verb (or intended sense).

This leads to several errors (e.g.):

\***um**kehre for kehre **um** 'I turn around'

# Aspect in Russian

Russian inflection is conditioned by an inherent feature known as aspect. For instance, the "perfective" verb *sorvat'* 'to pick' has a synthetic future, whereas the closely-related "imperfective" *sryvat'* forms an analytic future (using future-tense forms of the verb *byt'* 'to be'). This results in several errors (e.g.):

*\*budeš'* *sorvat'* for *sorvëš'* (2sg. fut.)

# Vowel harmony in Finnish compounds

The first stem of a noun compound does not participate in suffix harmony (Hakulinen et al. 2008:§18). For instance, the partitive singular of *lapinsirri* 'Temminck's stint' (a type of bird) is *lapinsirriä* because this is a compound of *Lapin* 'Laplandish' and *sirri* 'stint', and because all vowels in the second stem are neutral. Yet we have errors of the form (e.g.):

*lapinsirria* for *lapinsirriä*

# Allomorphy error patterns

We have seemingly rediscovered what linguists have long known: some allomorphy is genuinely unpredictable, including

- inherent features not included in the UniMorph feature bundles, like animacy and aspect in Slavic, and
- highly abstract morphophonological patterns, like Germanic ablaut and umlaut, Finnish consonant gradation, Hungarian linking vowels, Slavic yers, and Spanish diphthongization.

# A connection

Gouskova and Becker (2013) and Becker and Gouskova (2016) develop computational models of yer-deletion in Russian...

...and Albright et al. (2001) and Albright (2003) develop computational models of Spanish diphthongization...

...and both explicitly reject the type of morphophonological abstractness we find even modern neural networks need.

But, their evaluations are limited to speakers' nonce word judgments!



# Spelling errors

These are relatively rare overall:

- Dutch: \**geupgraded* for *geüpgraded* 'upgraded'
- English: \**disentered* for *disenterred*
- German: errors with proper use of *s*, *ss*, and *ß*
- Spanish: \**fungo* for *funjo* 'I service as'
- Portuguese: \**influisse* for *inflúisse* 'I would influence'

# Is automation the future?

One might prefer to automate error classification so that

- the system could be integrated into a rapid development process, or
- used as an additional objective during model training

so long as it has reasonably high agreement with human experts. Ideally such a system would scale to arbitrary languages, not just those for which we were able to quickly find linguistic expertise.

We pursued this, in a somewhat different vein, in the SIGMORPHON 2021 task on grapheme-to-phoneme conversion (Ashby et al. 2021).

Q. Di, E. Vylomova, and T. Baldwin. 2019. Modelling Tibetan verbal morphology. In *Proceedings of ALTA*, in press. ALTA. Sydney.

Elsner, M., Sims, A. D., Erdmann, A., Hernandez, A., Jaffe, A., Jin, L., ... , and Stevens, Guille, S. 2019. Modeling morphological learning, typology, and change: what can the neural sequence-to-sequence framework contribute? *Journal of Language Modeling* 7(1): 125-170.

[your name here?]