# Weird Inflects but OK: Making Sense of Morphological Generation Errors

**Kyle Gorman**[*], **Arya D. McCarthy**[†], **Ryan Cotterell**[†],
**Ekaterina Vylomova**[‡], **Miikka Silfverberg**[§], and **Magdalena Markowska**[*]
[*]The Graduate Center, City University of New York, [†]Johns Hopkins University,
[‡]University of Melbourne, [§]University of Helsinki
kgorman@gc.cuny.edu, arya@jhu.edu, ryan.cotterell@jhu.edu,
evylomova@gmail.com, miikka.silfverberg@helsinki.fi,
mmarkowska@gradcenter.cuny.edu

## Abstract

We conduct a manual error analysis of the CoNLL–SIGMORPHON 2017 Shared Task on Morphological Reinflection. In this task, systems are given a word in citation form (e.g., *hug*) and asked to produce the corresponding inflected form (e.g., the simple past *hugged*). This design lets us analyze errors much like we might analyze children's production errors. We propose an error taxonomy and use it to annotate errors made by the top two systems across twelve languages. Many of the observed errors are related to inflectional patterns sensitive to inherent linguistic properties such as animacy or affect; many others are failures to predict truly unpredictable inflectional behaviors. We also find nearly one quarter of the residual "errors" reflect errors in the gold data.

## 1 Introduction

A huge amount of work in natural language processing treats words as indivisible units, but the vast majority of the world's languages have rich word-internal structure. For instance, 80% of the languages analyzed in the *World Atlas of Linguistic Structure* (Dryer and Haspelmath 2013) inflect verbs for tense and 65% inflect nouns for case. Generating and processing complex words is thus crucial for multilingual speech and language technologies.

Recent work on large-scale, multilingual computational modeling of morphology (e.g., Durrett and DeNero 2013, Cotterell et al. 2016) targets supervised **inflection generation**. Such tasks require variable-length outputs, so they are less constrained than earlier segmentation-based tasks (e.g., Kurimo et al. 2010), but appear to be tractable with existing neural network–based models. For example, in the CoNLL–SIGMORPHON 2017 Shared Task (sub-task 1 and the "high-data condition"), the focus of this study, the best-ranked system generates novel inflectional forms with 90% accuracy or better for 46 out of the 52 target languages, It achieves perfect accuracy for four languages (Cotterell et al. 2017).

In light of these apparent success, we examine the failure modes of existing models for morphological generation. We first propose and motivate an error taxonomy for this task, inspired by similar proposals for other natural language generation and processing technologies such as grammatical error correction (e.g., Rozovskaya and Roth 2016) and machine translation (e.g., Popović and Ney 2011, Fishel et al. 2012, Irvine et al. 2013). We then use this taxonomy to perform a manual error analysis of the CoNLL–SIGMORPHON 2017 Shared Task. Such analyses can help to identify strengths and weaknesses of existing systems, suggest future improvements, and guide development of strong ensemble models, but are often neglected or treated as an afterthought. This annotation also allows us to measure the quality of the gold data.

Generating morphologically complex forms is a skill typically-developing children effortlessly acquire, so this task, and systems' error patterns, may have implications for the theory of language acquisition. While the shared task training paradigm is quite unlike human language learning, inference and evaluation resemble the classic *wug*-test (Berko 1958), in which speakers are presented with a word—either real or nonce—in citation form and prompted to provide a particular inflectional form of that word. Therefore, one can analyze inflection generation errors much like how one might analyze errors made by a child acquiring their first language. And, one can ask whether humans' and artificial learners' errors are in any way alike.

To answer these questions, we examine errors made by the two top-performing systems in the CoNLL–SIGMORPHON 2017 Shared Task for twelve languages.



*Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 140–151
Hong Kong, China, November 3-4, 2019. 

## 2 Materials and methods

Here, we describe the shared task, data sources, and the targeted systems.

### 2.1 The task

The CoNLL–SIGMORPHON 2017 Shared Task (Cotterell et al. 2017) consists of two supervised morphological generation sub-tasks across 52 languages. In sub-task 1, the training data consists of triples of lemma, inflectional bundle, and inflected form, as in Table 1. At inference time, the system is given lemmata and inflectional bundles and asked to produce the appropriate inflected forms. In sub-task 2, training data consists of complete inflectional paradigms, and at inference time, the system is asked to produce full paradigms for unseen lemmata. We focus on the results from sub-task 1, primarily because only two of the twelve teams chose to compete in sub-task 2. However, the proposed error taxonomy could easily be applied to sub-task 2, or to later morphological generation challenges such as sub-task 2 of the CoNLL–SIGMORPHON 2018 shared task (Cotterell et al. 2018) or sub-task 1 of the SIGMORPHON 2019 shared task (McCarthy et al. 2019).

#### 2.1.1 Data

The data in both sub-tasks is primarily sampled from UniMorph (Kirov et al. 2016, 2018), a free morphological database. In turn, UniMorph data for our twelve languages is automatically extracted from Wiktionary, a collaborative multilingual online dictionary. UniMorph pairs the cells of Wiktionary morphological paradigms, which bear prose labels like "genitive plural", to feature bundles in a language-independent morphological schema (Sylak-Glassman et al. 2015; also see Sylak-Glassman 2016). The data consist of the aforementioned triples of lemma, inflectional bundle, and inflected form. For sub-task 1, these triples were sampled from UniMorph paradigms according to frequencies of inflected forms as estimated from Wikipedia. Because of this sampling procedure, the data is sparse in the sense that there are rarely more than a few inflected forms per lemma. As such, this roughly mimics the statistical properties of the primary linguistic data encountered by child language learners (e.g., Chan 2008:71–100). Systems were evaluated under three training data conditions: low (100 triples), medium (1,000 triples) and high (10,000 triples). We focus on the high-data condition because nearly all systems performed poorly in the low- and medium-data conditions.

### 2.2 Systems

In sub-task 1, systems were ranked using the macro-averaged "per form" (i.e., full-token match) accuracy across all 52 target languages.[1] We analyze errors made by the two top-ranked systems, briefly described below.

UE-LMU-I **(Bergmanis et al. 2017)** This system uses a recurrent neural network (RNN) with a bidirectional gated recurrent unit (GRU) encoder, a unidirectional GRU decoder, and a standard attention mechanism. It enhances a closely-related competitor system (Kann and Schütze 2017) by augmenting the provided training data with identical input-output pairs so as to create a bias toward copying the input stem. It is ranked as the best-performing system on sub-task 1 (macro-average accuracy 95.32%).

CLUZH-7 **(Makarov et al. 2017)** This system also uses a neural encoder-decoder but replaces the "soft" attention mechanism with hard monotonic attention (Aharoni and Goldberg 2017) and special edit operations. It is ranked second-best overall on sub-task 2 (macro-average accuracy 95.12%) and also achieves the highest per form accuracy on eight languages including Hungarian and Spanish.

## 3 Error taxonomy

One major distinction in the proposed taxonomy of inflection generation errors is between those errors which can be given a linguistic characterization—i.e., in terms of misapplication of inflectional patterns independently attested in the target language—from those which cannot. As such we are inspired by a long and contentious debate in computational morphology research. Rumelhart and McClelland (1986) propose an early neural network model trained to generate the phonological form of English simple past tense verbs given the present tense. They claim that under in certain conditions, their model produces errors that are similar to those made by children acquiring English, such as **catched* for *caught*.[2]

[1] The other metric used in sub-task 1, average Levenshtein distance between word and target averaged over languages, ranks systems nearly identically (Cotterell et al. 2017:11).

[2] Such errors are known as **overregularizations** in the language acquisition literature (e.g., Marcus et al. 1992).

| Language | Lemma | Inflection | Inflected form |
|---|---|---|---|
| English | hug | V;PST | hugged |
| | spark | V;V.PTCP;PRS | sparking |
| German | aufbauen | V;IND;PRS;2;SG | baust auf |
| | Ärtzin | N;DAT;PL | Ärtzinnen |
| Spanish | descomponer | V;NEG;IMP;2;PL | no descompagáis |
| | liberar | V;IND;FUT;2;SG | liberarás |

Table 1: Sample training data for sub-task of the CoNLL–SIGMORPHON 2017 Shared Task. Each training example maps a **lemma** (a citation form) and **inflection** (a bundle of UniMorph morphosyntactic features) to an **inflected form**. At inference time, the inflected form is predicted given a lemma and inflection.

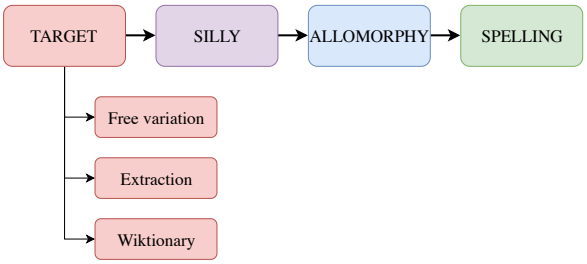


Figure 1: Overview of our annotation scheme, including subcategories. Annotators are instructed to proceed through the taxonomy from left to right.

Pinker and Prince (1988) and Sproat (1992:216f.) dispute this characterization, pointing out bizarre errors like **membled* for *mailed*. More recently, Kirov and Cotterell (2018) claim that modern neural network architectures—such as those used in the CoNLL–SIGMORPHON 2017 Shared Task—generalize reasonably well while largely eliminating these bizarre errors. However, Corkery et al. (2019) argue that the Kirov and Cotterell model predictions align poorly with human productions, and suggest that the reported results may be uncharacteristic due to fortuitous random seeding.

We desired a somewhat richer set of errors than this prior work. The final taxonomy—incorporating feedback from a ten-language pilot study—consists of four major error categories, with several additional sub-categories. The categories are applied sequentially, as in Figure 1. We now describe these categories.

**Target errors** This category consists of cases where the gold data is incorrect or incomplete.[3] We discern three sub-categories of target errors.

**Free variation errors** occur when more than one acceptable inflected form exists, but only one is present in the UniMorph data. **Extraction errors** indicate flaws in UniMorph's parsing of Wiktionary inflectional paradigms. **Wiktionary errors** represent errors in the Wiktionary data itself.

**Silly errors** This category consists of those "bizarre" errors which defy any purely linguistic characterization. In addition to the aforementioned case of **membled*, such errors have also been reported for other language generation tasks such as machine translation (Arthur et al. 2016) and text normalization (Gorman and Sproat 2016, Sproat and Jaitly 2017, Zhang et al. 2019).

**Allomorphy errors** This category consists of those errors which are characterized by misapplication of existing (i.e., independently attested) allomorphic patterns in the target language. Our annotation scheme recognizes four sub-categories of allomorphy error, but we set aside their their description for reasons of space.

**Spelling errors** This category includes inflected forms that do not follow language-specific orthographic conventions but are otherwise correct.

## 4 Results

We performed full error annotation on twelve of the 52 languages. Several other languages were initially targeted for annotation but produced too few errors to draw meaningful conclusions. Annotations were performed by the authors, all specialists in computational linguistics.[4] Of these, four languages—English, Finnish, Polish, and Russian—were annotated by native

[3] This label is applied regardless of whether the predicted inflected form is correct or not, and therefore is independent of system predictions. Furthermore, it is possible that both the gold data and prediction have the same incorrect inflected form, but detecting such cases is challenging.

[4] We do not claim that this level of expertise is strictly necessary; it might be the case that linguistically naïve native speakers could be trained to produce reliable annotations.

speakers; the remaining eight were annotated by second-language speakers. In addition to the annotation guidelines, annotators were encouraged to consult authoritative dictionaries and reference grammars—such as the *Iso suomen kielioppi* (Hakulinen et al. 2008) for Finnish, the *Duden* for German, the *Oxford Latin Dictionary* (Lee 1968), or the *Diccionario de la lengua española* for Spanish—and native speakers. Table 2 reports summary statistics for fully-annotated languages.

### 4.1 Inter-annotator agreement

Table 3 provides raw agreement and Krippendorf's $\alpha$ (Artstein and Poesio 2008) for those languages known to two annotators. As mentioned above, each annotator is a specialist in computational linguistics, and annotated at least one other language as well. Raw agreement is high, and while chance-corrected agreement statistics like $\alpha$ are notoriously difficult to interpret, $\alpha \geq 0.8$, a threshold obtained for all three double-coded languages, is generally considered to indicate substantial reliability (Krippendorff 2004:241f.).

### 4.2 Errors

Table 4 provides the counts of the four major categories of error for all twelve languages and for both systems. We now proceed to describe some patterns observed within these categories.

#### 4.2.1 Target errors

Table 5 gives counts for the three sub-categories of target errors.[5]

**Free variation errors** Finnish has several free variation errors, many involving vowel harmony. For example, the abessive suffix has two allomorphs, namely the the back-harmonic *-tta* and the front-harmonic *-ttä*. The lemma *progestiini* 'progestin' can take the back allomorph, giving *progestiini**tta***, but, vowel harmony often fails to apply when there are many intervening neutral vowels (*i* and *e*) between the harmonic trigger and the suffix (Hakulinen et al. 2008:§17), as is the case here. Therefore, the form *progestiini**ttä***, predicted by both systems, is grammatical, though not the form given by UniMorph. Another type of free variation error affects allomorphs of the Finnish genitive plural (gen.pl.). For instance, *omeno**iden***, *omeno**itten***, *omeno**jen***, *omen**ien*** and *omena**in*** are all possible gen.pl. forms of *omena* 'apple', but only one is present in UniMorph.

**Extraction errors** The comparatively low accuracy on Hungarian—CLUZH-7, the best performing system on this language, achieves 89.80% per form accuracy—appears to be due to large number of extraction errors. In most cases, the error comes from pairing one paradigm cell with another cell's inflectional bundle. For instance, UniMorph incorrectly labels **lagúnák* as the accusative plural (acc.pl.) for *lagúna* 'lagoon'; it is in fact the nominative plural (nom.pl.). In Romanian, a header for the Wiktionary paradigms reading "definite articulation" is incorrectly taken as an inflected form itself! Latin also suffers from pervasive extraction errors. This language has a robust phonemic contrast between short and long monophthongs (e.g., *malus* 'unpleasant' vs. *mālus* 'apple tree'). Long monophthongs are—at least in modern editions—indicated by the macron, a horizontal line above the vowel. UniMorph extraction has somehow removed macrons from all lemmata, though they are still present in the inflected forms. Thus, systems must attempt to predict an unpredictable phonemic contrast while mapping from lemma to inflected form. As a result, the vast majority of Latin errors concern vowel length.

**Wiktionary errors** Errors in the Wiktionary data itself are relatively rare and largely non-systematic. For example, in Spanish, **demarce* is given as the first person singular (1sg.) present subjunctive of *demarcar* 'to demarcate' instead of the correct form *demarque*.

#### 4.2.2 Silly errors

Silly errors were found for all languages except English; they also appear to be somewhat more common for UE-LMU-I (59) than for CLUZH-7 (37). UE-LMU-I predicts *praes**ōs*** as the acc.pl. of the Latin noun *praesul* (a title used by Roman religious leaders); there is no obvious analogue for the *ul-ōs* stem change. In German, CLUZH-7 unexpectedly truncates the gen.pl. form of the compound noun *Schädlingsbekämpfungsmittel* 'pesticide' to produce **Schädlingsbekämpfungsmit*. For the dative plural (dat.pl.) of the Russian compound noun meaning 'forced labor', UE-LMU-I inexplicably deletes the *r* of *rabóty* 'labor', giving the bizarre **prinudítel'nym **abótam***.[6] And, in Spanish,

[5] Some analyses conducted by Richard Sproat (p.c.) suggest that sub-task 2 was also highly affected by target errors.

[6] In the shared task, Russian data is given in the standard Cyrillic orthography; we have taken the liberty of romanizing

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

Table 2: Raw error counts out of 1,000 test examples for the target languages. Checkmarks indicate whether UniMorph data was available for a given major category in that language. Error overlap is the percentage of errors made by both systems. There were 1,701 errors in total (823 from UE-LMU-I and 878 from CLUZH-7).

| Language | RA | $\alpha$ |
|---|---|---|
| Dutch | 0.949 | 0.907 |
| English | 0.861 | 0.855 |
| Spanish | 0.861 | 0.875 |

Table 3: Inter-annotator agreement statistics for three double-coded languages. RA: raw agreement.

UE-LMU-I gives **at**ue**ngáis* as the second person plural present subjunctive of *atener* 'to maintain'. There is no analogue for this *e-ue* stem change.

#### 4.2.3 Allomorphy errors

With the exception of Hungarian and Latin—which suffer from systematic extraction errors—allomorphy errors are the largest category of error in all languages.

**Stem-final vowels in Finnish** In Finnish nouns and adjectives, stem-final vowels commonly disappear or alternate with *e* or *o* when the plural marker *i* is added to the stem (Hakulinen et al. 2008:§45). For instance, the inessive plural of *las**i*** 'glass' is *las**e**issa*. In principle, such alternations are predictable given the syllable count of the nominal stem, the stem-final consonants and the penultimate vowel, though the exact conditions are rather complex (Hakulinen et al. 2008:§46–50). For the compound noun *pohjanpystykorva* 'norrbottenspets' (a breed of dog), CLUZH-7 predicts an incorrect gen.pl. form **pohjanpystykorv**o**jen* for intended *pohjanpystykorvien*; it has transformed the stem final *a* to *o* and then selected the wrong plural marker (**-jen* instead of *-ien*) as a result.

**Ablaut in Dutch and German** Stem vowel alternations in the Germanic strong verbs are known as ablaut. Ablaut is robust in Dutch and German, and in both languages, is occasionally misapplied. In Dutch, for example, both systems overapply ablaut to the 1sg. preterite indicative forms of *printen* 'to print', producing **pr**o**nt* instead of *pr**i**ntte*. Similar errors are found in German. Both systems underapply ablaut to the 1sg. preterite indicative form of *saufen* 'to drink', producing **s**au**fte* instead of the expected *s**o**ff*, and UE-LMU-I overapplies ablaut to the third person preterite subjunctive of the weak verb *versenken* 'to sink', giving **vers**ä**chten* in place of the expected *vers**e**nkten*.

**Umlaut in German** Another stem change seen in German inflection is umlaut, which converts a *u*, *o*, or *a* in the final syllable of a stem changes to the corresponding front vowel *ü*, *ö*, or *ä*, respectively. Umlaut applies in many different morphological contexts (Hieble 1957), but most saliently in many plural nouns. One or both systems underapply umlaut in otherwise-correct plural forms of *Aasvogel* 'carrion bird', *Augenarzt* 'eye doctor', *Brunst* 'arousal', *Chalkogenidglas* 'chalcogenide glass', *Dachschaden* 'mental issues' (lit. 'roof damage'), *Energiezustand* 'energy level', *Hang* 'slope', *Stiefvater* 'stepfather', *Tibetfuchs* 'Tibetan fox', and *Vertrag* 'treaty'. But the systems also overapply umlaut in **Einwohnerzähle* (from

it here so as to make the data accessible to a wider audience.

| Language | Target | Silly | | Allomorphy | | Spelling | |
|---|---|---|---|---|---|---|---|
| | | UE-LMU-I | CLUZH-7 | UE-LMU-I | CLUZH-7 | UE-LMU-I | CLUZH-7 |
| Dutch | 8 | 1 | 1 | 19 | 16 | 5 | 7 |
| English | 3 | 0 | 0 | 18 | 18 | 7 | 11 |
| Finnish | 11 | 7 | 7 | 33 | 48 | 0 | 0 |
| German | 3 | 4 | 10 | 54 | 67 | 9 | 9 |
| Hungarian | 83 | 21 | 9 | 37 | 44 | 1 | 0 |
| Italian | 5 | 5 | 1 | 11 | 16 | 0 | 2 |
| Latin | 119 | 2 | 0 | 76 | 93 | 0 | 0 |
| Polish | 5 | 6 | 3 | 60 | 67 | 2 | 4 |
| Portuguese | 1 | 1 | 0 | 6 | 7 | 1 | 2 |
| Romanian | 54 | 3 | 5 | 61 | 69 | 1 | 2 |
| Russian | 7 | 7 | 0 | 48 | 45 | 23 | 28 |
| Spanish | 7 | 2 | 1 | 12 | 12 | 6 | 6 |
| Total | 306 | 59 | 37 | 435 | 502 | 55 | 71 |

Table 4: Error type counts by language and system; target error counts are combined across the two systems.

| Language | FV | Extraction | Wiktionary |
|---|---|---|---|
| Dutch | 0 | 3 | 5 |
| English | 0 | 2 | 1 |
| Finnish | 7 | 2 | 2 |
| German | 0 | 0 | 3 |
| Hungarian | 0 | 83 | 0 |
| Italian | 0 | 0 | 5 |
| Latin | 0 | 118 | 1 |
| Polish | 0 | 4 | 1 |
| Portuguese | 0 | 1 | 0 |
| Romanian | 1 | 51 | 2 |
| Russian | 1 | 5 | 1 |
| Spanish | 2 | 3 | 2 |
| Total | 11 | 272 | 23 |

Table 5: A breakdown of target errors by sub-category; counts are combined across the two systems. FV: free variation errors; Extraction: extraction errors; Wiktionary: Wiktionary errors.

*Einwohnerzahl* 'population', **Förmer* (from *Form* 'shape'), **Neuwähle* (from *Neuwahl* 're-vote'), and **Sprösse* (from *Spross* 'bud').

**Consonant gradation in Finnish** Many Finnish words undergo a set of unpredictable stem changes known as consonant gradation. Here, a "strong grade" of a consonant–normally a voiceless stop like *t*—alternates with the weak grade—a voiced stop like *d*—but the stop may also delete in the weak grade (Hakulinen et al. 2008:§41). Gradation leads to inflection errors because not all lexemes participate in gradation, and because the weak grade of the stem consonant is not predictable from the lemma. For instance, CLUZH-7 incorrectly applies the weak grade to the negated third person singular **ei kiemur**d**a* (from *kiemurtaa* 'to crawl'); the proper gradation is *t-r* instead of the predicted *t-d*. CLUZH-7 also incorrectly produces the strong grade where the weak grade is required, failing to delete the *k* in the comitative **rikosla**k**ein* (from *rikoslaki* 'criminal law').

**Linking vowels in Hungarian** The Hungarian noun plural suffix is *-k*, usually preceded by a *a*, *o*, *e*, or *ö* linking vowel. For example, the nom.pl. form of *vér* 'blood' is *vér**ek***. The choice of linking vowel is partly determined by vowel harmony: back vowel stems select *a* or *o* whereas front vowel stems select *e* or *ö*. However, for back vowel stems, it is largely unpredictable whether *a* or *o* is used (Siptár and Törkenczy 2000:224f., Vago 1980:110f.), and there are several cases where one or both systems predict an incorrect linking vowel. For example, UE-LMU-I predicts an incorrect elative plural **masszázs**ak**ból* for *masszázs* 'massage'; the correct form is *masszázs**ok**ból*.

**Yers in Polish** Another sub-category of allomorphy error in Polish concerns the yers, the "fleeting vowels" of Slavic. Oblique forms of the Polish nouns *klęs**e**k* 'defeat' and *żag**ie**l* 'sail', for example lack a stem *e* or *ie*, respectively, in certain case forms, as seen in the gen.pl. *klęsk* and

*żagli*. Because fleeting vowels' position and quality are unpredictable, they cannot be analyzed as epenthetic. Instead, they are assumed to be present in the underlying form of certain roots and affixes, but somehow represented distinctly from the non-fleeting vowels (Lightner 1965, Rubach 1986). According to the analysis, a yer is deleted except when the following syllable also contains an yer, and the fleeting *e* and *ie* surface in the nom.sg. forms above because the masculine nom.sg. suffix is itself a yer (Gussman 1980:36f., Rubach 1984:41). It is impossible to predict the position or quality of a yer without referring to the rest of the inflectional paradigm,[7] and this indeterminacy contributes to several inflectional errors. For instance, CLUZH-7 predicts **żagieli* instead of the expected *żagli*, and both systems predict **klęsek* for of the expected *klęsk*. Similar errors are also found in Russian.

**Spanish diphthongization** Many Spanish verbs exhibit a stem change in which mid vowels *e* and *o* in the final syllable of the stem diphthongize to *ie* [je] and *ue* [we], respectively, when they bear primary stress. Whether or not a mid vowel participates in diphthongization is largely unpredictable (Brame and Bordelois 1974:132f., Harris 1969:74f.).[8] For example, *negar* 'to deny' undergoes diphthongization (e.g., 1sg. present indicative ***nie***go), but *pegar* 'to stick' does not (1sg. present indicative *pego*). Both models underapply diphthongization in **despl**e**gue* (from *desplegar* 'to unfold') and **rec**o**la* (from *recolar* 'to strain again'). Interestingly, these are 1st conjugation (i.e., *-ar*) verbs, and children acquiring Spanish tend to underapply diphthongization in this class (Mayol 2007). But CLUZH-7 also overapplies diphthongization in **atañ**ie**res* (from *atañer* 'to concern') and **gañ**ie**seis* (from *gañir* 'to yelp'). Similar errors occur in Portuguese, which also exhibits a stress-conditioned stem vowel alternation.

**Noun plural suffixes in German** German has five major noun plural suffixes, and many errors involve the use of the wrong plural. The most frequent pattern is the overapplication of the *-(e)n* plural—traditionally regarded as the most productive plural suffix (Bech 1963, Wunderlich 1999)—as in *Eosin* 'eosin', *Fußballweltmeisterschaftsqualifikationsspiel* 'football world championship qualification game', *Hartung*, a poetic term for 'January', *Karbonatit* 'carbonatite', *Metallatom* 'metal atom', and *Vorjahr* 'last year'. Overapplication of *-e* is also common, as in *Abonnement* 'subscription', *Etat* 'budget', *Funke* 'spark', *Katholic* 'Catholic', *Königsgelb* 'yellow pigment', *Reaktorbau* 'reactor construction', *Prinzess* 'princess', *Toupet* 'toupee'. Interestingly, both types of error are produced by children acquiring German (e.g., Clahsen 1999, Marcus et al. 1995, Szagun 2001).

**Genitive singular suffixes in Polish** Polish has two gen.sg. suffixes, *-a* and *-u*. It is generally impossible to predict which gen.sg. allomorph a given stem will select, and there is no evidence that one is more productive than the other (Dąbrowska 2001, 2005, Kottum 1981, Maunsch 2003). This unpredictable allomorphy causes many gen.sg. errors to both systems, such as **ateusz**u*** for *ateusz**a*** 'atheist', **izotop**a*** for *izotop**u*** 'isotope', **krzyk**a*** for *krzyk**u*** 'scream', and **legar**u*** for *legar**a*** 'joist'.

**Verbal prefixes in German** Some verbal prefixes in German are known as "separable" because they separate (i.e., are postposed) from their host verb when tensed. Others, the "inseparable" prefixes, are always attached to their host verb without exception. Finally, some prefixes, such as *um-*, are separable or inseparable depending on the verb, and this leads to several errors. For example, both systems predict **__um__kehre* for the 1sg. present indicative of *umkehren* 'to turn around'; the correct form is the separable *kehre __um__*.

**Animacy in Polish and Russian** Case syncretisms in inanimate (i.e., non-personal) nouns are found in many Slavic languages. However, animacy is an inherent feature of nouns and cannot be predicted from the form of the lemma alone. In Russian, for example, CLUZH-7 wrongly predicts a syncretic acc.pl. for the animate *sadist* 'id.' and both systems incorrectly predict a distinct (i.e., non-syncretic) acc.sg. for the inanimate *magazin* 'shop'. Similarly in Polish, both systems predict incorrect syncretic accusatives for animates such as *śpiewak* 'singer' and *Żyd* 'Jew', and incorrect non-syncretic accusatives for inanimates such as *szampan* 'champagne'. Some Polish stem changes are also conditioned by animacy. For example, for the inanimate noun *katalizator* 'catalyst', both sys-

[7] Gouskova and Becker (2013) and Becker and Gouskova (2016) develop formal models of yer-deletion in Russian, but do not evaluate performance on actual held-out words.

[8] Albright et al. (2001) and Albright (2003) develop a computational model to predict Spanish diphthongization, but do not report its performance on actual held-out verb forms.

tems incorrectly predict a nom.pl. **katalizatorzy* instead of *katalizatory*; the mutation of *r* to *rz* before the nom.pl. *-y* is restricted to masculine animates (Feldstein 2001:27).

**Aspect in Russian** Russian verbal inflection is conditioned by an inherent feature known as aspect. For instance, the perfective verb *sorvat'* 'to pick' forms a synthetic future whereas the closely-related imperfective *sryvat'* forms a periphrastic (i.e., multi-word) future formed using future-tense forms of *byt'* 'to be'. Several errors involve the wrong future form for a verb's aspect. For example, for the perfective *sorvat'*, CLUZH-7 incorrectly predicts a periphrastic second person singular future **budeš' sorvat'* instead of the expected synthetic *sorvëš'*.

**Vowel harmony in Finnish compounds** In Finnish, the first stem in a noun compound does not participate in suffix harmony (Hakulinen et al. 2008:§14). For example, the partitive singular of the compound *lapinsirri* 'Temminck's stint' (a type of bird) is the *lapinsirriä*. Because this lemma is a compound of *Lapin* 'of Lapland' and *sirri* 'stint', and because all vowels in the second second stem of the compound are neutral, front harmony—the default—applies. However, CLUZH-7 generates **lapinsirria*, a form which would be correct were the lemma not a compound.

**Internal inflection in Russian compounds** Many Russian nouns in the shared task are adjective-noun or noun-noun compounds, and systems fail to appropriately inflect both components of the compound. The acc.pl. of *lëgkaja promyšlennost'* 'light industry' is *lëgkie promyšlennosti*, but UE-LMU-I predicts **lëgkix promyšlennosti*, incorrectly placing the adjective in the genitive case. Other adjective-noun compounds for which one or both of the systems fail to produce proper agreement morphology include *vizitnaja kartočka* 'business card' and *bulevo množestvo* 'boolean domain'. Both stems of most noun-noun compounds, particularly hyphenated compounds, are inflected. For example, the prepositional plural of *gosudarstvo-donor* 'donor state' is *gosudarstvax-donorax*, but both systems predict **gosudarstvo-donorax*, in which only the second stem is inflected. However, there are some cases in which one stem of a compound is not declined. For instance, in *sindrom Aspergera* 'Asperger's syndrome', only the head noun *sindrom* should be inflected because *Aspergera* is a nominal modifier and already in genitive case, but both systems incorrectly inflect the second stem producing the gen.pl. **sindromov Asperger*.

#### 4.2.4 Spelling errors

Spelling errors are relatively rare overall. In Dutch, diaeresis is used to mark hiatus—adjacent vowels in consecutive syllables—and thus the past participle of *upgraden* 'to upgrade' should be *geüpgraded*, not the predicted **geupgraded*. Several English errors concern an orthographic doubling of certain final consonants; for example, both systems predict a past participle **disentered* instead of the expected *disenterred*. There are many German spelling errors, including several concerning the spelling of the gen.sg. suffix—written as *-es* or *-s* depending on context—or *s*, *ss*, and *ß*, all pronounced [s]. In Spanish, a *g* followed by *i* or *e* is read as [x], not as [g], so the verb *fungir* 'to service as' has a 1sg. future indicative spelled *funjo* rather than the predicted **fungo*. Several Portuguese and Spanish predictions omit the acute accent used to indicate exceptional primary stress; e.g., Portuguese **influisse* for the 1sg. imperfect subjunctive *influísse* (from *influir* 'to influence').

## 5 Discussion

### 5.1 Target errors

Target errors heavily impact performance for Hungarian, Latin, and Romanian. Overall, nearly one fourth of our sample's errors were target errors, and we suspect such errors also lurk in the training and development data. Clearly, the UniMorph data used in this task requires further vetting.

### 5.2 Allomorphy errors

Overall, silly errors were far less common than allomorphy errors. Many of the allomorphy errors appear to result from unpredictable linguistic behaviors rather than failures to extract reliable generalizations. In some cases, errors reflect systems' inability to predict inherent features such as animacy and aspect in Slavic. Such features are not encoded in UniMorph, although this information is often present on Wiktionary. Generally speaking, these features cannot be predicted from the orthographic form of lemmata,[9] but we

[9] Certain prefixes and stress patterns are cues to aspect in Russian verbs (Wade 2010:268), but this is not true of inherent features in general.

suspect that the relevant information could be induced using either contextual or type-level word embeddings. We leave this for future work.[10] Systems also appear to struggle with lemmata which are themselves internally complex due to word-formation processes like prefixation or compounding, including prefix verbs in German and compounds in Finnish and Russian. Lemma-internal structure, once again, is not currently encoded in UniMorph, though it could in principle be extracted from Wiktionary entries. Finally, we see that systems struggle with certain lexically-specific morphophonological patterns—Germanic ablaut and umlaut, Finnish consonant gradation, Hungarian linking vowels, Slavic yers, and Spanish diphthongization—and with lexically-conditioned affix selection in German and Polish. We have seemingly rediscovered what linguists have long known: certain allomorphic patterns cannot be predicted from the form of lemmata alone; they must be memorized. It is unreasonable to expect any neural network, no matter how powerful, to predict what is truly unpredictable.

Our analysis is limited to languages included the shared task, those for which the top systems have a non-trivial number of errors, and those for which we have sufficient linguistic expertise. As a result, our final sample of twelve languages only includes two major language families, Indo-European and Uralic, the latter represented by Finnish and Hungarian. However, this sample has some degree of grammatical diversity. Linguists traditionally distinguish between two types of morphological exponence. In **agglutination**, each morphological feature corresponds roughly to a single affix. For instance, in the Hungarian form *cinkosoknak*, the dat.pl. of *cinkos* 'accomplice', the *-ok* suffix marks plurality and the *-nak* suffix indicates the dative case. In **fusion**, on the other hand, single affixes may realize many morphological features at once. For instance, in the Russian form *čabrecov*, the gen.pl. of *čabrec* 'thyme', the *-ov* suffix is both genitive and plural (and its form also indirectly indicates that the stem is masculine). Agglutination is characteristic of the Uralic languages, whereas Indo-European languages makes heavy use of fusion. Furthermore, vowel harmony is limited to the two Uralic languages.

[10] Sub-task two of the SIGMORPHON 2019 Shared Task (McCarthy et al. 2019) involves lemmatization and morphological analysis in sentential context, but the applicability of this to the inflection task has not yet received much attention.

## 6 Conclusion

We propose an error taxonomy for morphological inflection generation and apply it to the predictions of the two best systems in the CoNLL–SIGMORPHON 2017 Shared Task. We estimate a lower bound for the percentage of "target" errors in the gold data. Over 80% of the remaining (non-target) errors can be understood as misapplication of language-specific morphological or spelling principles. One potential remedy is to enrich the input linguistic representations with, e.g., compound structure and inherent grammatical features; however, this is unlikely to avoid all errors; some morphological patterns cannot be generalized but only memorized.

The above analysis depends on manual annotation, but one might prefer to automate error classification. An automated system, for example, could be integrated into a rapid development process, or used as an additional objective during training and tuning, 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 linguistic expertise is readily available. A powerful ensemble model could help identify candidate target errors, and for certain high-resource languages, it might be possible to leverage finite-state morphological analyzers and lexicons to distinguish between silly, spelling, and allomorphy errors. We leave these and many other open questions for future work.

## Acknowledgments

Ekaterina Levitskaya made substantial contributions to early stages of this project. Suzanne van der Feest assisted with Dutch annotations, and Alëna Aksënova assisted with Russian annotations. Richard Sproat provided spiritual guidance and contributed an impressionistic characterization of the sub-task 2 error profiles.

Miikka Silfverberg was funded in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 771113).

## References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*,

pages 2004–2015, Vancouver. Association for Computational Linguistics.

Adam Albright. 2003. A quantitative study of Spanish paradigm gaps. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, pages 1–14, Somerville, MA. Cascadilla.

Adam Albright, Argelia Andrade, and Bruce Hayes. 2001. Segmental environments of Spanish diphthongization. *UCLA Working Papers in Linguistics*, 7:117–151.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Gunnar Bech. 1963. Zur Morphologie der deutschen Substantive. *Lingua*, 12(1):177–189.

Michael Becker and Maria Gouskova. 2016. Source-oriented generalizations as grammar inference in Russian vowel deletion. *Linguistic Inquiry*, 47(3):391–425.

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, Vancouver. Association for Computational Linguistics.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14:150–177.

Michael K. Brame and Ivonne Bordelois. 1974. Some controversial questions in Spanish phonology. *Linguistic Inquiry*, 5(2):282–298.

Erwin Chan. 2008. *Structures and distributions in morphology learning*. Ph.D. thesis, University of Pennsylvania.

Harald Clahsen. 1999. Lexical entries and rules of language: a multidisciplinary study of German inflection. *Behavioral and Brain Sciences*, 22(6):991–1069.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL–SIGMORPHON 2017 shared task: universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. `https://wals.info/`.

Dudenredaktion, editor. No date. *Duden in 12 Banden. 1: Die deutsche Rechtschreibung*, 27th edition. Bibliographisches Institut, Berlin.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta. Association for Computational Linguistics.

Ewa Dąbrowska. 2001. Learning a morphological system without a default: the Polish genitive. *Journal of Child Language*, 28(3):545–574.

Ewa Dąbrowska. 2005. Productivity and beyond: mastering the Polish genitive. *Journal of Child Language*, 32:191–205.

Ronald F. Feldstein. 2001. *A concise Polish grammar*. Slavic and East European Language Resource Center, Durham, NC.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. TerrorCat: a translation error categorization-based MT quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, Montréal. Association for Computational Linguistics.

Kyle Gorman and Richard Sproat. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics*, 4:507–519.

Maria Gouskova and Michael Becker. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory*, 31(3):735–765.

Edmund Gussman. 1980. *Studies in abstract phonology*. MIT Press, Cambridge.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. 2008. *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura, Helsinki.

James Harris. 1969. *Spanish phonology*. MIT Press, Cambridge.

Jacob Hieble. 1957. What about the German umlaut? *The German Quarterly*, 30(4):272–274.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Katharina Kann and Hinrich Schütze. 2017. The LMU system for the CoNLL–SIGMORPHON 2017 shared task on universal morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 40–48, Vancouver. Association for Computational Linguistics.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: universal morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 1868–1873, Miyazaki. European Language Resource Association.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 3121–3126, Portorož. European Language Resources Association.

Steiner E. Kottum. 1981. The genitive singular form of masculine nouns in Polish. *Scando-Slavica*, 27(1):179–186.

Klaus Krippendorff. 2004. *Content analysis: an introduction to its methodology*, 2nd edition. Sage, Thousand Oaks, CA.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. Morpho Challenge 2005–2010: evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95, Uppsala. Association for Computational Linguistics.

G. M. Lee, editor. 1968. *Oxford Latin dictionary*. Claredon Press, Oxford.

Theodore M. Lightner. 1965. *Segmental phonology of Modern Standard Russian*. Ph.D. thesis, MIT.

Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 49–57, Vancouver. Association for Computational Linguistics.

Gary Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: the exception that proves the rule. *Cognitive Psychology*, 29(3):189–256.

Gary Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, John Rosen, and Fei Xu. 1992. *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

Hanna Maunsch. 2003. Current alternations in inflection of Polish masculine inanimate nouns in the singular: a pilot study. *Investigationes Linguisticae*, 9:4–21.

Laia Mayol. 2007. Acquisition of irregular patterns in Spanish verbal morphology. In *12th ESSLLI Student Session Proceedings*, pages 185–196, Dublin. Association for Logic, Language and Information.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence. Association for Computational Linguistics.

Steven Pinker and Alan Prince. 1988. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1–2):73–193.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Real Academia Española, editor. 1992. *Diccionario de la lengua española*, 21st edition. Real Academia Española, Madrid.

Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2205–2215, Berlin. Association for Computational Linguistics.

Jerzy Rubach. 1984. *Cyclic and lexical phonology: the structure of Polish*. Foris, Dordrecht.

Jerzy Rubach. 1986. Abstract vowels in three-dimensional phonology: the yers. *The Linguistic Review*, 5(3):247–280.

David Rumelhart and Jay McClelland. 1986. On learning the past tenses of English verbs. In Jay McClelland, David Rumelhart, and the PDP Research Group, editors, *Parallel distributed processing: explorations into the microstructure of cognition. Vol. 2: Psychological and biological models*, pages 216–271. Bradford Books, Cambridge.

Péter Siptár and Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford University Press, Oxford.

Richard Sproat. 1992. *Morphology and computation*. MIT Press, Cambridge.

Richard Sproat and Navdeep Jaitly. 2017. An RNN model of text normalization. In *Proceedings of Interspeech 2017*, pages 754–758, Stockholm. International Speech Communication Association.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (UniMorph schema). Technical report, Department of Computer Science, Johns Hopkins University.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 674–680, Beijing. Association for Computational Linguistics.

Gisela Szagun. 2001. Learning different regularities: the acquisition of noun plurals by German-speaking children. *First Language*, 21:109–141.

Robert M. Vago. 1980. *The sound pattern of Hungarian*. Georgetown University Press, Washington, D.C.

Terence Wade. 2010. *A comprehensive Russian grammar*, 2nd edition. Wiley-Blackwell, Oxford.

Dieter Wunderlich. 1999. German noun plural reconsidered. *Behavioral & Brain Science*, 22(6):1044–1045.

Hao Zhang, Richard Sproat, Axel H. Ng., Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.