SIGMORPHON–UniMorph 2022 Shared Task 0: Modeling Inflection in Language Acquisition

Jordan Kodner and **Salam Khalifa**
Stony Brook University
Department of Linguistics and Institute for Advanced Computational Science
Stony Brook, NY USA
{jordan.kodner,salam.khalifa}@stonybrook.edu

Abstract

This year's iteration of the SIGMORPHON-UniMorph shared task on "human-like" morphological inflection generation focuses on generalization and errors in language acquisition. Systems are trained on data sets extracted from corpora of child-directed speech in order to simulate a natural learning setting, and their predictions are evaluated against what is known about children's developmental trajectories for three well-studied patterns: English past tense, German noun plurals, and Arabic noun plurals. Three submitted neural systems were evaluated together with two baselines. Performance was generally good, and all systems were prone to human-like over-regularization. However, all systems were also prone to non-human-like over-irregularization and nonsense productions to varying degrees. We situate this behavior in a discussion of the Past Tense Debate.[1]

1 Introduction

The overarching goal of this subtask of the 2022 SIGMORPHON-UniMorph shared task on morphological inflection, in contrast with this year's and previous years' typologically informed subtasks, was to provide insight into how current state-of-the-art morphological inflection models relate to human language acquirers, to what extent they behave similarly or differently, and in what respects they perform better or worse. As such, the task was designed to be cognitively informative while still approachable for the NLP morphology community. This was achieved in two ways: First, nested training sets of increasing size were extracted from corpora of child-directed speech, following (Belth et al., 2021), allow us to approximate learning trajectories with batch learning models that are typical in the field today rather than incremental models which might better approximate the child language acquisition setting. Second, supervision with semantic features substitutes for semantic information which children in real acquisition settings would certainly glean from their linguistic and environmental experiences. While this simplified the task considerably, it also permitted us to focus on the act of generating correct forms in the absence of other learning confounds.

1.1 Historical Background

The acquisition of morphological patterns has been heavily investigated for decades from both experimental and computational perspectives. The acquisition of English past tense in particular was the original locus of the so-called "Past Tense Debate," with implications not only for the nature of cognitive morphological representations (*single-route* or *dual-route*), but also for the nature of cognitive representations and computations more generally (symbolic or non-symbolic, distributed or not). The debate kicked off in earnest following the publication of an early connectionist (psychologically-inspired feed-forward artificial neural network) model for past tense learning (Rumelhart and McClelland, 1986). The model did not explicitly handle regular and irregular patterns differently (it was single-route), yet it performed reasonably well given the computing power and neural network know-how available at the time.

A response by Pinker and Prince (Pinker and Prince, 1988), who instead advocated for a symbolic model of past tense learning and representation in which regular and irregular forms were handled separately (a dual-route model) was the first in what turned into many years and dozens of papers worth of discussion. As the years passed, they expanded to encompass morphological patterns in other languages as well, particularly pluralization of German nouns. See McClelland and Patterson (2002) and Pinker and Ullman (2002) for surveys of the debate.

[1] Data, evaluation scripts, and predictions are available at: https://github.com/sigmorphon/2022InflectionST

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 157 - 175
July 14, 2022 ©2022 Association for Computational Linguistics

Modern deep neural systems are in many ways the spiritual and technological successors to the connectionists. Given the success of such models on a wide range of tasks in NLP, it is possible that modern neural morphology models could overcome many of the drawbacks of their predecessors. A recent paper (Kirov and Cotterell, 2018) made this argument to the computational linguistics community. Given the critical responses and responses to the responses so far (Corkery et al., 2019; McCurdy et al., 2020; Belth et al., 2021; Beser, 2021; Dankers et al., 2021), it is fair to say that the debate has been reignited.

1.2 Contribution of the Shared Task

The SIGMORPHON inflection shared task paradigm (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021) is well-suited for assessing the behavior of morphological learning systems. Developing a greater understanding of the ways in which systems are or are not human-like can help explain why their prediction accuracy is so good and also direct us towards areas of improvement. The Past Tense Debate and the developmental research that came out of it provides a backdrop over which we can evaluate the systems.

This is the second year that the SIGMORPHON shared task on morphological inflection is running a "human-like" generalization task. Last year's task[2] investigated the extent to which computational systems matched adult acceptability ratings on *wug tests* presented in English, German, Dutch and Russian. Such a task is suited for testing systems' ability to form human-like analogies between phonologically related forms in a laboratory setting. However, the task is not suitable for answering the questions addressed this year.

Adults appear to approach wug tests differently from children (Schütze, 2005), with many adults treating it as a game that requires clever analogies (Derwing and Baker, 1977). This difference is observed in the original wug test study (Berko, 1958), in which adults readily produced analogical past forms *glung* and *glang* for *gling* on analogy with verbs like *sting-stung* and *sing-sung*, while 83 of 86 young children either produced *glinged* or refused to answer. It is not clear to what extent this is a difference in child and adult linguistic representations or an artifact of experimental design. It is also not entirely clear to what extent gradient acceptability ratings are the result of the gradient experimental prompts, since they may drive test subjects to spread responses over a wider range than they would otherwise (Parducci and Perrett, 1971).[3] See Yang (2020) for additional discussion.

Since this task sought to compare computational morphology learning systems to child learners, we took a different approach. Teams were asked to train inflection models as for previous SIGMORPHON shared tasks but on data drawn from corpora of child directed speech, the input that children receive during acquisition. Systems made predictions on real words rather than nonce words, simulating the experience of children who need to produce never before heard forms for lemmas that they already know. These outputs were compared to what is known about children's learning trajectories and errorful productions.

Three inflectional patterns, English past tense, German noun plurals, and Arabic noun plurals, were chosen because they have been heavily studied from a developmental perspective and have been subject to computational cognitive modeling research. The acquisition of English past tense and German noun pluralization in particular have received renewed interest in recent years, and while less work has been conducted on this aspect of Arabic, we believe that it will make for an elucidating challenge case going forward. The remainder of this section briefly summarizes some relevant findings for English, German, and Arabic.

1.3 English Past Tense

The general state of the English past tense system is a familiar one. There is a clearly productive general default *regular* suffix *-ed* (subject to phonologically-conditioned allomorphy) which applies to the vast majority of verbs and new coinings, as well as several much less frequent patterns usually described as *irregular*. Many of these irregulars indicate past tense through a stem vowel mutation (the so-called *strong verbs* paralleled in other Germanic languages), for example, *sing-sang*, *sting-stung*, *bite-bit*, and *ride-rode*. Others combine a stem mutation with a coronal suffix (the so-called *semi-weak* verbs, where regular *-ed* verbs

[2] 2021 description and data available here: https://github.com/sigmorphon/2021Task0

[3] Armstrong et al. (1983) presents a stark example of this, finding that participants would gradiently rate integers for their "evenness" given the opportunity, even though the even/odd distinction is completely binary.

are *weak*) including *keep-kept*, *sleep-slept* and *tell-told*. There are also a few one-off suppletive forms, most notably *go-went*.

There is a clear distinction to be made between the single overwhelming majority default pattern, and the rest. Nevertheless, the irregulars as a whole tend to fall in the high end of the frequency range and so are over-represented in the input. As a result, children identify *-ed* as productive later than one may expect given its high type frequency. They acquire it around age three (Berko, 1958; Marcus et al., 1992). It is hard to say exactly what verbal vocabulary size this age corresponds to since there is quite a lot of variation among individuals, but Marcus et al. (1992, ch. 5) report that Sarah and Adam from the Brown Corpus (Brown et al., 1973) have produced 300-350 unique verbs by age three.

Children's novel productions exhibit an asymmetry between *over-regularizations*, which are over-applications of the default pattern (e.g., **goed*, **feeled*) and *over-irregularizations*, which apply irregular patterns to regular verbs (e.g., *fry-*frew* by analogy with *fly-flew* or *peep-*pept* by analogy with *keep-kept* and *sleep-slept*).

The former error type is far more common than the latter, both in English and in other languages. Studies of past tense errors in English learners have found over-irregularization rates of under 0.2% (Xu and Pinker, 1995), but over-regularization rates orders of magnitiude higher between 8 and 10% (Maratsos, 2000; Yang, 2002; Maslen et al., 2004). Similar findings have been observed in German past participle production with under 1% over-irregularization and about 10% over-regularization (Clahsen and Rothweiler, 1993), and a similar ratio in Spanish verbal production (Clahsen et al., 1992; Mayol, 2007). See Marcus et al. (1992) for more discussion. Nevertheless, for all their strengths, over-irregularization has been a persistent challenge for single-route models since the early connectionist days. Early connectionist models were also prone to producing nonsense, for example *mail-membled* (Xu and Pinker, 1995).

Despite its mundanity, the English past tense system provides a valuable test case for models of morphology acquisition. That said, it does have a major drawback. Since there is only one apparently productive global default pattern, and that pattern applies to the overwhelming majority of types, a naive model that performs simple frequency matching is expected to perform quite well on English.

Corpus	-*e*%	-*(e)n*%	-*er*%	-∅%	-*s*%
CELEX	27	48	4	17	4
UniMorph	34.4	37.3	2.9	19.2	4.0

Table 1: Type distribution of German noun plural types in CELEX (Baayen et al., 1993) reported in Sonnenstuhl and Huth (2002), and in UniMorph as reported in McCurdy et al. (2020). 2.1% of UniMorph nouns have "other" plural forms.

While type frequency is certainly the most important factor in the acquisition of productive generalizations (Aronoff, 1976; MacWhinney, 1978; Bybee, 1985; Baayen, 1993; Elman, 1998; Pierrehumbert, 2003; Yang, 2016), this obscures potential differences between dramatically different learning models. German noun pluralization was introduced into the Past Tense Debate because it has a much more even distribution of inflectional patterns.

1.4 German Noun Plurals

Unlike English past tense, the German noun plural system has several relatively frequent pluralization patterns: -*(e)n*, -*e*, -*er*, -∅ and -*s* with distributions summarized in Table 1. Pluralization may be further indicated with Umlaut, or the fronting of certain vowels. There are three Umlaut patterns which are clearly indicated in German orthography: (*a*→*ä*, *o*→*ö*, *u*→*ü*). Suffixing and Umlaut appear to be largely orthogonal, so some recent computational modeling work has focused only on the former (McCurdy et al., 2020; Belth et al., 2021).

It is clear that German noun plurals do not have a high-frequency global default like English. However, some plural forms appear to be defaults for nouns that meet certain conditions. Feminine nouns, for example, productively pluralize with -*(e)n*, where the vowel is subject to phonologically conditioned allomorphy (Wiese, 1996). Several phonotactic properties are also shown to correlate with pluralization type preferences (Zaretsky and Lange, 2015).

While the -*s* plural is the least frequent of the language's pluralization types, it has attracted considerable theoretical attention because it nevertheless appears to be a case of a minority default pattern (Clahsen, 1990; Marcus et al., 1995; Sonnenstuhl and Huth, 2002). The -*s* plural is the plural of last resort that speakers fall back on when the conditions for other plurals are not met, however, unlike English -*ed*, it is not particularly frequent. As a result, it serves as a means of differentiating learning models which rely naively on type frequency from ones which leverage type frequency to learn more underlyingly complex morphological systems.

Developmental studies show that children do successfully learn this system around the same age that English past tense is acquired. Children learn *-e* *-∅*, and *-(e)n* by the time they know 100 words, and while *-er* and *-s* are learned later, they are acquired reliably around 500 words (Elsen, 2002). Over-application of *-(e)n* is the most common error type followed by over-application of *-e*, though even *-s* and *-er* are overproduced (Elsen, 2002).

1.5 Arabic Noun Plurals

Finally, we introduce Arabic noun pluralization as another challenge case. Arabic nouns may form plurals in two ways: by suffixation (so-called *sound plurals*) or by stem mutation (so-called *broken plurals*). There are two sound plural suffixes, a feminine *-āt*, and a masculine *-ūn* (*-īn, -ū, or -ī* depending on a nominal's case and state). The relationship between gender and sound plural ending is reliable but not exceptionless. In particular, some masculine nouns, generally non-human masculine nouns, take the feminine sound plural, e.g., *imtiḥān-imtiḥān-āt* 'exam.' Noun gender can be determined with agreement – pronouns, adjectives, and verbs all agree with nouns in gender, so masculine nouns taking feminine plurals are a clear morphological mismatch.

Broken plurals can be divided into many subclasses by which templatic pattern defines the output of their stem mutations. In Modern Standard Arabic (MSA), there are approximately 30 broken plural patterns (McCarthy and Prince, 1990), though the exact count depends on the level of abstraction assumed for the templatic pattern. Some classes of singular templates are known to take specific plural patterns, e.g., *maktab* (maCCaC) 'desk, office' → *makātib* (maCāCiC). On the other hand, different singular patterns can take the same plural pattern, e.g., both *kitāb* (CiCāC) 'book' and *sarīr* (CaCīC) 'bed' are pluralized as *kutub* and *surur* (CuCuC), respectively. This results in a very complex system. There are many theoretical accounts which seek to explain and predict the mappings betwen singular and broken plural patterns. McCarthy and Prince (1990), for example, group the broken plural patterns according to prosodic shapes and concluded that the iambic pattern is a productive one. However, some of their findings have been challenged (Gaskell and Marslen-Wilson, 2001; Haddad, 2008).

The Arabic pluralization system is quite elaborate, and it is not completely acquired by children until primary school age, however, most properties of the system are acquired much earlier, in line with the timelines observed for English and German (Ravid and Farah, 1999). Using a wug test paradigm, Ravid and Farah (1999) demonstrate that children follow *u*-shaped learning trajectories due to transient over-regularization in the direction *broken* → *sound*, and over-regularization in the direction MASC *sound* → FEM *sound*. The vast majority of child production errors belong to one of these two types, an asymmetry consistent with strong tendency for over-regularization rather than over-irregularization observed for other languages.

Dawdy-Hesterberg and Pierrehumbert (2014) present a series of related exemplar learning models and apply them to Arabic data. Their systems are generally successful at learning Arabic plural patterns, but they show fewer MASC *sound* → FEM *sound* and far more *sound* → *broken* errors than are observed in children. Exemplar learners are a kind of single-route learner, so this lack of asymmetry in error types may be expected given what has been observed for English.

2 Task Description

This task was organized very similarly to other iterations of the inflection task from the participants' perspective in order to encourage cross-submissions with this year's large scale generalization inflection task (Kodner et al., 2022). Participants were asked to design supervised learning systems which could predict an inflected form given a lemma and a morphological feature set corresponding to an inflectional category or cell in a morphological paradigm. They were provided with several nested training sets as well as a development set and test set for each language. The train and dev sets consisted of (`lemma`, `inflected`, `feature set`) triples, while the inflected forms were held out from the test set.

Initially, only training and development sets were available to participants. They were expected to design, train, and tune their models on this data. Shortly before the submission deadline, test sets with held-out inflections were released. In contrast with the large-scale subtask and previous iterations, only three languages were investigated which could be evaluated in detail: American English, Standard German, and Modern Standard Arabic. Several nested training sets were released for each language in increments of 100 items. Participants were asked to return predictions from models trained on each training size.

3 Data Preparation

Data sets for (American) English and (Standard) German were extracted from the CHILDES collection of child-directed speech (CDS) corpora (MacWhinney, 2000). CHILDES contains several types of corpora with various types of annotation. English was sourced from the Brown (Brown et al., 1973)[4] and Brent (Brent and Siskind, 2001) corpora. These contain free dialogue between caregivers and their children alternating with lines of morphological annotation. In (1), `*MOT` indicates that this utterance was produced by the child's mother and `%mor` indicates that the following line contains POS tags, lemmas, and morphological features. However, "words" in morphological annotation lines do not consistently line up one-to-one with tokens in dialogue lines, so it is not feasible to match lemma-feature pairs to inflected forms. To accomplish this, features were converted into UniMorph format, and (`lemma`, `inflected`, `features`) triples were extracted from English UniMorph (McCarthy et al., 2020).

(1) **Adam `021016.cha` 571-572** (Brown, 1973)

```
*MOT: what are you writing ?
%mor: pro:int|what aux|be&PRES
  pro:per|you part|write-PRESP ?
```

One advantage of CHILDES is that it presents vocabulary that a typical child is likely exposed to during the acquisition process, and since it contains dialogue, it can also be used to make reasonable frequency estimates of child-directed speech. In NLP terms, it is a reasonable approximation of the training set over which children learn morphological inflection. See Kodner (2022) for more information. 2,054 nouns were sampled from CHILDES weighted by their CHILDES frequencies, and their plurals were extracted from UniMorph. 454 of these items were sampled uniformly and reserved as the development set. 600 of the remainder were uniformly sampled from the remainder and set aside as the test set. The remaining 1,000 was used as the maximum training set. Smaller nested subsets in increments of 100 were sampled from these, weighted by noun lemma frequencies in CHILDES such that each larger subset was a superset of the smaller.

Training and test were sampled uniformly with respect to one another to guarantee that the test set would contain interesting test items. Another reasonable approach would have been to sample the 1,000 training items by frequency from the entire data set and then sample the test items from the remainder in order to yield a training set containing more frequent items and a test set containing less frequent items. Since item token frequency correlates with age of acquisition (Goodman et al., 2008), this would correspond to a realistic scenario where systems predict later-acquired forms from their knowledge of earlier acquired forms. However, English past tense irregulars (i.e., non-*ed* pasts), are heavily skewed towards the high end of a Zipfian frequency distribution, so such an approach would not yield many interesting test items.

The German data set was created in much the same way as the English with CDS frequency information sourced from the CHILDES Leo corpus (Behrens, 2006) and nominative plural forms matched from German UniMorph. Gender is known to be a predictor for plural forms (Wiese, 1996), so the German UniMorph features were augmented with MASC, FEM, or NEUT gender tags converted from the CHILDES annotation lines. These were split into 600 training items, 500 development items, and 600 test items with the same frequency-weighted algorithm that was applied to English. The intersection of nouns extracted from Leo and nouns present in UniMorph was relatively small, so the largest training set that could be extracted only contains 600 items.

Ideally, the Arabic data set would also be extracted from a CDS corpus in order to get a reasonable estimation of a child's vocabulary. Colloquial Arabic varieties are unfortunately considered to be low-resourced in terms of available linguistic resources, so even though there are several dialectal CDS corpora (Kern et al., 2009; Alqattan, 2015; Salama and Alansary, 2017), they do not provide morphological annotations useful to the task in hand. Thus, we selected Modern Standard Arabic (MSA) for the shared task. Even though it has virtually no native speakers and no CDS corpora, it is

[4] This is a classic CDS corpus built by Roger Brown. It is not to be confused with the classic NLP Brown Corpus developed at Brown University (Kučera and Francis, 1967).

well-resourced and exhibits the same kinds of morphological patterns present across Arabic varieties. A reasonable workaround for the lack of CDS is to estimate a child size corpus from a given non-CDS corpus through lemma frequencies. This will most likely contain high frequency lexemes that typically do not appear in CDS corpora but will likely cover a similar distribution of morphological phenomena (Kodner, 2019).

For this shared task, the Arabic data set was sourced from the Penn Arabic Treebank (PATB) (Maamouri et al., 2004), which is a morphologically and syntactically annotated news corpus of MSA. The corpus is written using standard Arabic orthography and it is fully diacritized. Diacritization include short vowels, specific case and state markings, and gemination. Arabic text without diacritization does not mark these critical phonological segments and thus would not be useful for the task at hand. Despite including fine-grained morphological annotations, PATB lacks the annotations of functional (grammatical) gender and number in addition to rationality (animacy). Therefore, a version that has been enriched with additional features through the CALIMA$_{MSA}$ morphological analyzer (Taji et al., 2018) was used. Plural inflections that reflect state and case were normalized to a single inflection since only pluralization was under investigation for this task.

The 2,000 most frequent plural nouns were extracted according to their lemma frequencies from the TRAIN split of PATB (Diab et al., 2013). These were then split into a training set of 1,000 items, a development set of 343 items, and a test set of 600 items using the same algorithm that split English and German. An animacy feature HUM or NON-HUM was added was added to each noun, since it is known to impact nominal inflection patterns (McCarthy and Prince, 1990).

4 Systems

The same neural and non-neural baselines were provided for this task and the 2022 typologically diverse inflection shared task. The neural system `Neural`, Wu et al. (2021), is a character-level transformer. It is identical to the system `CHR-TRM` which was used in the 2021 task with identical hyperparameters. The non-neural system, `NonNeur`, is identical to the non-neural baseline made available in 2021 and 2020.[5] Three systems were submitted, the first and last of which were also submitted to the large scale generalization task:

`CLUZH` (Silvan Wehrli and Makarov, 2022): Universität Zürich's system is identical to the one submitted to this year's large scale generalization subtask (Kodner et al., 2022). Their submission is a character-level transducer which operates over edit actions: insertion, deletion, substitution, and copy. They implement true mini-batch training for a substantial speed up, rendering the system more practical on larger training sets.

`HeiMorph` (Ramarao et al., 2022): The team from Heinrich-Heine-Universität Düsseldorf developed a system with a self-attention Transformer architecture with bigram hallucination. Submitted models were trained on the enriched data setsthat include either 1,000 or 10,000 bigram-aware hallucinated word pairs, generated separately for each training set size. The system was implemented with Fairseq, a Pytorch-based tool.

`OSU` (Elsner and Court, 2022): OSU's system is identical to the one submitted to this year's large scale generalization subtask. This inflection system is a transformer whose input is augmented with an analogical exemplar showing how to inflect a different word into the target cell. In addition, alignment-based heuristic features indicate how well the exemplar is likely to match the output.

5 Evaluation

Whole-form accuracy was employed as the primary quantitative evaluation, though several further analyses were carried out by partitioning data over grammatical gender and other factors. Performance was good overall but showed some points of divergence from human behavior. This section provides an analysis for each of the shared task's three languages.

5.1 English Past Tense

As expected given its majority default pattern, performance across all systems was higher on English than the other languages. Table 2 summarizes the results. `CLUZH` in particular achieved most of its performance already on 100 training items, while `HeiMorph` and the neural baseline show the most substantial gains as the training size increases.

[5] Available here: https://github.com/sigmorphon/2022InflectionST/tree/main/baselines/nonneural

#Train	CLUZH	HeiM	OSU	*Neural*	*NonN*
Avg.	**85.67**	65.65	81.48	70.12	80.60
100	**80.33**	50.50	67.67	21.67	68.17
200	**82.33**	68.17	75.00	46.83	75.67
300	**83.17**	64.83	78.50	62.83	77.50
400	**83.50**	46.17	81.67	72.83	80.00
500	**85.67**	69.17	81.67	78.17	81.17
600	**87.83**	69.17	83.50	82.33	83.17
700	**87.00**	69.33	85.00	84.00	84.00
800	**87.83**	70.33	85.17	83.17	84.33
900	**90.33**	71.50	88.00	84.50	85.50
1000	**88.67**	77.33	**88.67**	84.83	**86.50**
Ortho	**91.17**	82.0	90.67		

Table 2: English: Overall percent exact match training size for submitted systems and baselines. *Ortho* are accuracy at 1000 when stem-final spelling errors are not penalized.

CLUZH	Match	Other	Reg	*-ed*	?
100	4.65	4.65	88.37	88.37	2.33
500	9.3	6.98	83.72	83.72	0.0
1000	9.3	6.98	83.72	83.72	0.0

HeiM	Match	Other	Reg	*-ed*	?
100	9.3	18.6	58.14	69.77	2.33
500	6.98	37.21	46.51	51.16	4.65
1000	2.33	9.3	76.74	81.4	6.98

OSU	Match	Other	Reg	*-ed*	?
100	9.3	27.91	53.49	55.81	6.98
500	11.63	9.3	67.44	74.42	4.65
1000	2.33	4.65	88.37	90.7	2.33

Table 3: Error type analysis for English irregular verbs. *Match* = % correct. *Other* = % other plausible strong and weak irregulars. *Reg* = % "correct" regularized. *-ed* = % forms ending in -ed. *?* = other nonsense output

Since English orthography is notoriously complex, evaluating this task on written English presents an unnecessary additional burden on the systems. And though few errors could be clearly attributed to orthography in practice, some were found. In particular, some systems occasionally failed to follow orthographic rules regarding the doubling of word-final consonants. For example, systems produced **enthraled* instead of expected *enthralled* and **payed* for *paid*. These are spelling mistakes, though the latter is actually attested in Early Modern English texts. The final line in Table 2, *Ortho*, evaluates the submitted systems at 1,000 training when these particular errors are not penalized.

The performance of each system rises 2-5 points when these errors are ignored. There is, however, one cause for concern. 557 of 600 test items form regular *-ed* pasts, so a baseline system which always predicts *-ed* should achieve 92.83% accuracy in the *Ortho* evaluation. No system outperformed this baseline.

Table 3 investigates the role that over-regularization played in driving errors at 100, 500, and 1,000 training. Numbers for other training sizes are available in Table 15 in the Appendix. The *Match* column presents the percent of gold irregulars which were correctly predicted. These values are appropriately low given that these patterns are generally unpredictable in English. The *Other* column indicates the percent of gold irregulars which were subject to other plausible irregular patterns (e.g., OSU produced *bring-?brang*, which is incorrect according to the gold standard *brought*[6]). The sum of these two columns is the proportion of gold irregulars that were predicted to be irregular. HeiMorph and OSU produced substantially more irregular forms than CLUZH.

Columns *Reg* and *-ed* indicate the rate of over-regularization. *Reg* is the proportion of gold irregular items that were inflected as "correct" regular past forms (e.g., *buy-*buyed*, *bleed-*bleeded*). This was the majority for each system at each training size, though CLUZH performed more over-regularization. The *-ed* adds predictions that included *-ed* but were still incorrect (e.g., *forgive-*forgaved* for expected *forgave*). *?* counts outputs that qualify as nonsense in some way (e.g., *seek-*sougk* for expected *sought*.)

Overall, the systems all clearly show a tendency towards over-regularization. The systems clearly learn an *-ed* rule and apply it readily. In fact, all the systems, especially CLUZH, are *too* good from a developmental perspective. They begin applying *-ed* the majority of the time after only 100 training instances, well ahead of children.

Table 4 and the full version Table 16 in the Appendix quantify over-irregularization. *Match* indicates percent of gold regular *-ed* verbs inflected correctly. *SorW* is the proportion gold regular verbs inflected according to some strong or semi-weak irregular pattern, for example OSU *ply-*plew*, CLUZH *spike-*spake*, and HeiMorph *top-*topt*. *SC+ed* is the proportion of gold regular verbs that received an *-ed* suffix but were also subjected to some stem vowel change (e.g., OSU *fine-*founed*), and *Irreg* is the sum total of irregularized gold regular verbs. *?* again indicates nonsense outputs

[6] This particular error is interesting. *Brang* does exist dialectally in American English.

CLUZH	Match	SorW	SC+*ed*	Irreg	?
100	99.1	0.9	0.0	0.9	0.0
500	97.49	2.51	0.0	2.51	0.0
1000	97.49	2.51	0.0	2.51	0.0

HeiM	Match	SorW	SC+*ed*	Irreg	?
100	63.91	12.93	0.72	14.9	21.18
500	80.43	15.44	0.72	16.88	2.69
1000	88.15	5.39	0.18	6.46	5.39

OSU	Match	SorW	SC+*ed*	Irreg	?
100	79.17	8.98	3.77	15.44	5.39
500	90.66	3.05	0.9	4.85	4.49
1000	97.49	1.26	0.36	1.62	0.9

Table 4: Error type analysis for English regular verbs. *Match* = % correct or orthographic. *SorW* = % well-formed strong or semi-weak irregular. *SC+ed* = % *-ed* is present but with a vowel change. *Irreg* = % all plausibly irregular patterns. *?* = nonsense output

	-ed	→*a*	→*u*	Other	?
Gold	2	2	3	1	–
CLUZH	4	1	3	0	0
HeiM	8	0	0	0	0
OSU	8	0	0	0	0

Table 5: Inflection type for English monosyllabic *-ing* verbs at 1,000 training. *-ed* = regular. →*a* = *sing-sang*-type. →*u* = *sting-stung*-type. *Other* = other inflection (*bring-brought* in the gold standard). *?* = nonsense inflection.

including *ski-*soa*, *crush-*crushi*, and *test-*tsot*.[7]

CLUZH produced by far the least over-irregularized forms at smaller training sizes, while the other systems produced substantially more. A qualitative error analysis revealed some interesting patterns. Every system extended the semi-weak shortening pattern of *keep-kept* to the lemmas such as *cheep* or *beep*, producing **bept* or **chept*. OSU and the neural baseline extended the *think-thought* pattern to monosyllabic verbs beginning with *consonant-h*, producing pairs such as *whiz-*whought* and *thin-*thought*. These are clear examples of unnatural over-irregularization behavior.

Finally, the monosyllabic *-ing* verbs were investigated as an illustrative study. Since it is not possible to predict the correct past forms of the *-ing* test items in a principled way, systems were expected to fail by raw accuracy. Thus, this makes for an interesting case for a more detailed analysis. There were six such items in the training data and eight in the test data. Lists of training and test items is provided in (2)-(3)[8] along with the smallest training sample in which the training items appeared.

(2) **Training**

```
300 swing-swung
300 sing-sang
700 thing-thinged
800 ding-dinged
800 sling-slung
900 cling-clung
```

(3) **Test**

```
sting-stung      bring-brought
fling-flung      king-kinged
ring-rang        spring-sprang
ping-pinged      string-strung
```

Even though the number of irregular *-ing* verbs increases with training size, over-regularization to *-ed* is the most common output at 1,000 training. HeiMorph and OSU "correctly" over-regularize all eight test items at 1,000 training. CLUZH over-regularizes half the forms and prefers *-ung* forms for three of the others (4). This ratio makes sense if the system is matching the training data, which has more *-ung* pasts than *-ang* pasts.

(4) **CLUZH *-ing* predictions at 1000 training**

```
sting-stung      bring-bringed
fling-flinged    king-kinged
ring-rang        spring-sprung
ping-pinged      string-strung
```

There was much more variety at smaller training sizes, including an aamusing incorrect production generated by the OSU system: it produced the present-past pair *ping-*pong*. Overall, systems showed a preference for over-regularization relative to over-irregularization, especially apparent for CLUZH. Nevertheless, they all produced orders of magnitude more over-irregularization than observed during child development as described in the Introduction. In particular, systems picked up on the semi-weak shortening pattern, over-applied *-ought*, and applied stem changes of various sorts even when simultaneously applying *-ed*. All systems showed super-human performance in their acquisition of *-ed*, productively applying it after only 100 training examples, when a human child might produce *-ed* only after learning a few hundred verbs (Brown, 1973).

[7] The neural baseline produced several instances of metathesis, especially at smaller training sizes. Examples include *flitter-*filtered*, *bark-*braked*, *sand-*snad*, *dodge-*dogde*, *clink-*clikned*, *own-*won*, *sell-*sleled*, *spring-*sprigned*, and *erase-*reased* at 100 vs. *erase-*earsed* at 200.

#Train	CLUZH	HeiM	OSU	*Neural*	*NonN*
Avg.	**76.72**	67.03	72.11	58.33	74.81
100	**72.67**	59.00	66.50	18.67	63.67
200	**74.67**	63.50	69.17	51.00	71.50
300	**76.17**	66.33	72.00	62.00	**76.00**
400	**78.17**	69.00	74.00	68.83	**78.00**
500	**78.50**	71.00	76.00	74.17	**79.50**
600	**80.17**	73.33	75.00	75.33	**80.17**
Suff.	**89.00**	85.83	85.67		
Uml.	**90.67**	88.83	90.17		

Table 6: German: Overall percent exact match training size for submitted systems and baselines. *Suff.* are accuracy at 600 when only suffix type is evaluated. *Uml.* are accuracy at 600 when only Umlaut is evaluated.

5.2 German Noun Pluralization

Performance on German, summarized in Table 6, was generally good but lower than for English at equivalent training sizes. This may be because German noun pluralization does not have an overwhelming majority pattern. CLUZH achieved the highest accuracies of any of the submitted systems, though it performed roughly on par with the non-neural baseline at training sizes 300 and above. All systems except for the neural baseline achieved most of their performance after only 100 training items – CLUZH in particular reached 90% of its final performance.

Two additional accuracy measures are reported in Table 6 for the submitted systems. *Suff* refers to test accuracy in the 600 training condition when only the suffix type is evaluated rather than exact match. This measure is more lenient because Umlaut and any other alternations do not need to be generated correctly. As expected, each system achieves a higher *Suff* score than exact match score at 600. HeiMorph shows the largest increase of 12.5 points. *Uml.* refers to test accuracy in the 600 training condition when only the presence of absence of Umlaut is evaluated. 522, or 87% of test items do not form plurals with additional Umlaut, so a baseline system that ignored the process altogether would achieve 87%. Each system surpassed this baseline by a small amount.

Table 7 presents Umlaut confusion matrices for each submitted system. Each system shows a similar pattern of under-application of Umlaut. Only HeiMorph applies Umlaut in more than half of the cases where it should apply, but only barely. Each system also occasionally over-applies Umlaut, with HeiMorph exhibiting the highest over-application rate at 5.75%.

CLUZH	Gold NC	Gold Umlaut
Pred NC	**506 (96.93%)**	40 (51.28%)
Pred Umlaut	16 (3.07%)	**38 (48.72%)**

HeiMorph	Gold NC	Gold Umlaut
Pred NC	492 (94.25%)	37 (47.44%)
Pred Umlaut	30 (5.75%)	41 (52.56%)

OSU	Gold NC	Gold Umlaut
Pred NC	503 (96.36%)	40 (51.28%)
Pred Umlaut	19 (3.64%)	**38 (48.72%)**

Table 7: German Umlaut/No Change confusion matrices at 600 training

Set	*-e%*	-(*e*)*n%*	*-er%*	-∅%	*-s%*	#
Train200	29.5	46.5	2.0	20.0	2.0	200
Train600	27.8	38.0	3.0	26.7	4.6	600
TrainF	2.8	96.2	0.0	0.5	0.5	212
TrainM	45.4	7.3	1.5	41.2	4.5	262
TrainN	33.3	4.0	11.1	40.5	11.1	126
Test	30.5	36.7	2.8	24.8	5.2	600
TestF	3.5	95.0	0.0	0.0	1.5	201
TestM	48.9	9.2	0.3	35.9	5.6	284
TestN	32.2	2.6	13.9	40.9	10.4	115

Table 8: Distribution of German plural suffixes in the 200 training set, and by gender in the 600 training and test sets.

Table 8 presents the overall and by-gender distribution of each pluralization suffix in the training and test sets. Counts for *-en* and *-n* are collapsed, since they are phonologically predictable allomorphs. These can be compared to the CELEX and UniMorph distributions presented in Table 1.

All systems are more accurate when the gold pluralization suffix is one of the three more common (*-e*, -(*e*)*n*, -∅) than one of the two less common (*-er*, *-s*). This is summarized in the confusion matrices provided in Tables 9-10 for training sizes 200 and 600. OSU and HeiMorph produces some forms containing miscellaneous stem-internal errors, marked as *?* in the confusion matrices, such as a *j* > *t* mutation in **Kabeltaue* as the plural of *Kabeljau*, but these were much rarer, and much more limited, than what was observed in their English predicitions. CLUZH did not produce any. *-er* and *-s* plurals were under-produced by each system. In both cases, each system usually applied *-e* instead. For example, CLUZH produced **Grase* instead of expected *Gräser* as the plural of *Gras*.

Comparing this to findings about the time course of children's plural pattern acquisition (Elsen, 2002), each system appears to acquire productive *-e* and -(*e*)*n* as early as expected, as evidenced by

[8] Some of these have alternative past forms in actual speech. Only a single form was chosen for each in the data set.

CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	**166**	17	17	2	27	229
P -(*e*)*n*	7	**198**	0	2	4	211
P -*er*	0	0	0	0	0	0
P -∅	10	5	0	**145**	0	160
P -*s*	0	0	0	0	0	0
P ?	0	0	0	0	0	0
Sum	183	220	17	149	31	600

HeiM	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	110	8	7	6	15	146
P -(*e*)*n*	22	192	0	5	6	225
P -*er*	3	0	1	1	2	7
P -∅	42	14	7	133	7	203
P -*s*	3	4	2	1	**1**	11
P ?	3	2	0	3	0	8
Sum	183	220	17	149	31	600

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	159	16	14	5	28	222
P -(*e*)*n*	10	183	0	0	2	195
P -*er*	0	2	**3**	0	0	5
P -∅	10	10	0	139	0	159
P -*s*	1	0	0	0	0	1
P ?	3	9	0	5	1	18
Sum	183	220	17	149	31	600

Table 9: German inflection confusion matrices at 200 training for FEM nouns only, disregarding Umlaut. *G* = Gold, *P* = Prediction.

CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	**168**	16	13	0	18	215
P -(*e*)*n*	6	**198**	0	1	2	207
P -*er*	0	0	3	0	0	3
P -∅	8	5	0	**148**	0	161
P -*s*	1	1	1	0	**11**	14
P ?	0	0	0	0	0	0
Sum	183	220	17	149	31	600

HeiM	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	154	13	12	4	16	199
P -(*e*)*n*	14	194	0	0	4	212
P -*er*	4	0	**4**	1	4	13
P -∅	9	10	0	142	1	162
P -*s*	1	1	1	0	3	6
P ?	1	2	0	2	3	8
Sum	183	220	17	149	31	600

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	155	19	13	1	18	206
P -(*e*)*n*	7	184	0	0	2	193
P -*er*	2	0	3	1	0	6
P -∅	11	10	1	142	1	165
P -*s*	2	1	0	1	8	12
P ?	6	6	0	4	2	18
Sum	183	220	17	149	31	600

Table 10: German inflection confusion matrices for each submitted system at 600 training disregarding Umlaut. *G* = Gold, *P* = Prediction.

over-application after 200 training. This is contrasted with -*er*, -*s*, which they rarely produce after 200 training but produce (still insufficiently frequently) at 600 training. These results are broadly consistent with what is observed developmentally, with the caveat that -*er*, -*s* are proportionately less frequent in the small training sets than the large ones (Table 8).

Since analyzing suffix confusions as a whole obscures some patterns, Tables 18-20 are provided in the Appendix which present confusion matrices partitioned by gender. Every system effectively learns that -(*e*)*n* is the appropriate ending for feminine nouns, and as observed in Table 18, most errors among feminines can be attributed to over-application of this ending.

Overall, systems show some consistency with the developmental patterns evaluated here. What the systems do learn, they learn on appropriate amounts of training data. However, they continue to greatly under-produce the infrequent but apparently minority default -*s* pattern. Further work needs to be done, along the lines of recent papers published on this topic (McCurdy et al., 2020; Belth et al., 2021; Dankers et al., 2021) to determine whether or not the submitted systems are behaving in a human-like manner.

5.3 Arabic Noun Pluralization

Arabic proved to be the most challenging of the three languages: summarized in Table 11, no system achieved more than 67% accuracy on any training size. This result is to be expected, since Arabic noun pluralization is more complex than the other phenomena evaluated. As for English, some errors were determined to be very minor and primarily orthographic. Not penalizing these errors yields the *Minor* line in the table, for which each system shows a 4-5-point increase. The line *S*F*S*M*B* additionally does not penalize broken-to-broken errors as long as the applied broken pattern is itself valid. This increases performance by another 6-9 points, indicating that predicting the correct broken pattern for an item was challenging compared to determining whether to apply a broken pattern at all. Since there are so many broken patterns, this is not surprising. Nevertheless, accuracies in this most permissive evaluation are still lower than for German or English.

Noun gender and rationality are known to correlate with plural formation in Arabic, so Table 12 presents the distribution of items by gender and rationality in the training and test sets. Masculine sound plurals are the least frequent, and masculine

#Train	CLUZH	HeiM	OSU	*Neural*	*NonN*
Avg.	**59.63**	55.37	57.53	52.70	33.70
100	**45.67**	41.83	34.00	14.83	28.33
200	**54.83**	45.67	49.17	41.67	28.33
300	**54.17**	48.67	53.33	51.00	29.00
400	**58.33**	49.83	54.17	52.83	31.67
500	**62.00**	59.67	61.00	57.17	34.83
600	63.17	62.83	**64.00**	61.50	35.50
700	**64.67**	60.33	63.83	62.50	36.33
800	**63.33**	62.17	63.83	61.33	37.33
900	64.33	63.33	**66.67**	60.83	37.33
1000	**65.83**	59.33	65.33	63.33	38.33
Minor	**69.67**	63.67	68.83		
SFSMB	75.50	71.00	**76.00**		

Table 11: Arabic: Overall percent exact match training size for submitted systems and baselines. *Minor* are accuracy at 1000 training when errors deemed to be minor or orthographic are ignored. *SFSMB* are accuracy at 1000 training when confusion between broken patterns is not penalized.

Set	SF	SM	B	#
Train	424	140	434	998
Train F	222	0	85	307
Train M	202	140	349	691
Train HUM	24	129	84	237
Train NHUM	400	11	350	761
Test	257	62	281	600
Test F	156	0	73	229
Test M	101	62	208	371
Test HUM	15	50	43	108
Test NHUM	242	12	238	492

Table 12: Distribution of Arabic plural types suffixes by gender and rationality in the 1000-training and test sets. Two irregular forms in the training set, *ðāt* 'self' and *ḥabb* 'seeds,' are excluded from this table.

nouns (as determined through agreement) are more diverse than feminines in their plural forms. About two thirds of feminine nouns take the feminine sound plural and all of the remainder take a broken plural. A plurality of rational nouns take the masculine sound plural, while non-rational nouns, which account for nearly five sixths of the data, are split about evenly between feminine sound and broken plurals with very few masculine sound plurals.

Table 13 presents confusion matrices for each plural type for each system. Breakdowns by gender and rationality can be found in Tables 22-25 in the Appendix. Each system over-produced feminine sound plurals at the expense of masculine sound and broken, but they varied in their production of masculine sound and broken plurals. This extended across gender and rationality.

Prior work evaluated children and a computational system according to their distributions of sound-to-sound, sound-to-broken, broken-to-sound, and broken-to-broken errors (Ravid and Farah, 1999; Dawdy-Hesterberg and Pierrehumbert, 2014). Table 14 provides such a breakdown for each system at 1,000 training, and Table 21 in the Appendix provides further breakdowns by gender and rationality. Each system's error types follow the same frequency order: broken-to-sound is the most frequent followed by broken-to-broken, sound-to-broken, and sound-to-sound errors.

CLUZH	Gold SF	Gold SM	Gold B	Sum
Pred SF	213	5	52	270
Pred SM	2	**51**	16	69
Pred B	38	4	**206**	248
Pred ?	4	2	7	13
Sum	257	62	281	600

HeiM	Gold SF	Gold SM	Gold B	Sum
Pred SF	**227**	7	72	306
Pred SM	3	43	15	61
Pred B	18	5	177	200
Pred ?	9	7	17	33
Sum	257	62	281	600

OSU	Gold SF	Gold SM	Gold B	Sum
Pred SF	218	8	49	275
Pred SM	5	50	15	70
Pred B	29	2	202	233
Pred ?	5	2	15	22
Sum	257	62	281	600

Table 13: Arabic inflection confusion matrices for each submitted system at 1000 training.

	S→S	S→B	B→S	B→B
CLUZH	7	42	68	52
HeiM	10	23	87	65
OSU	13	31	64	57

Table 14: Arabic error types at 1000 training.

This is quite unlike children, who overwhelmingly produce broken-to-sound and sound-to-sound errors (in both cases, mostly to feminine sound). It is also different from the (Dawdy-Hesterberg and Pierrehumbert, 2014) exemplar models in that broken-to-broken were much more common. Nevertheless, those exemplar models and the neural models submitted here both greatly over-produce sound-to-broken errors. The lack of to-broken errors among children, similar to the lack of over-irregularization in English, suggests that these are memorized patterns rather than ones that are productively applied. Thus, to-broken errors can be seen as a kind of over-irregularization.

6 Discussion

This year's shared task investigated the performance of neural systems on an inflection task designed to mimic language acquisition. Training data was mostly sourced from the CHILDES collection of child-directed speech corpora and extracted by frequency to represent early linguistic input, and systems produced past forms and plurals for real words, simulating children producing novel (to them) forms of lemmas that they know from daily life.

This was a challenging task characterized by small training data and complex patterns. Nevertheless, systems performed well in terms of raw accuracy. American English past tense forms proved the easiest, followed by Standard German noun plurals, then Modern Standard Arabic noun plurals. In some ways, the submitted systems actually outperformed children – they all learned the producive *-ed* pattern for English past tense after only 100 training items, far earlier than what is reported for children. Systems also achieved most of their performance on very small data. Superhuman performance on very small data is a valuable property for real-world NLP applications.

Compared to early connectionist systems, modern neural morphology learners produce far fewer nonsense forms of the *mail-membled* type, though this still remains a problem, even in the largest training conditions evaluated here. This is consistent with the findings of Gorman et al. (2019), which found that what they called "silly" errors were still present in the productions of the 2017 task, but they were majorly reduced compared to early work.

Systems "successfully" over-regularized the English *-ed* past, the most frequent German noun plural types, and the Arabic feminine sound plural. This is is a human-like tendency, however it cannot be said whether this indicates deep understanding of the paradigms or a simple case of frequency matching. Systems under-applied rarer German noun plural types even at the largest training size, which may imply the latter, but more work would need to be done to confirm this.

The most significant weakness of all three systems uncovered by this analysis is persistent inhuman over-irregularization. Though rates of over-irregularization varied significantly on English, all systems produced far more instances of it than child learners, and the problem was starker for Arabic. All three systems dramatically overproduced sound-to-broken and broken-to-broken errors which are rare in child productions. Broken plural patterns are apparently no more productive than English strong verb mutations, so their over-application has to be seen as over-irregularization.

Though Gorman et al. (2019) did not categorize errors in these cognitively-minded terms, they did find evidence for over-irregularization in their analysis. They noted, for example, that one system over-applied Spanish diphthongization, a pattern that applies to many verbs. The pattern is frequent but unpredictable – many verbs that could be subject to diphthongization are not. The pattern is apparently lexicalized and unproductive, and children under-apply it if anything (Mayol, 2007), thus the over-application is an instance of over-irregularization.

All of the systems evaluated this year happen to be neural *single-route* models that do not make an explicit distinction between regular and irregular items. No *dual-route* models were submitted for comparison. While all systems performed well, they showed the clear hallmarks of such models, in particular a tendency to over-produce over-irregularization. All of the technical improvements over the decades have greatly improved overall prediction accuracy, but single-route models are still single-route models.

What do these results tell us about human cognition? Even if the systems had shown very human-like performance, we could not therefore conclude that they are good models of cognition. As summarized recently in Guest and Martin (2021), that line of reasoning is backward. Prediction is not explanation. We would need to first justify the assertion that these are theoretically plausible cognitive models. Only then, if these systems were effective representations of cognition, then we should expect them behave in a human-like manner.

What studies like this do provide is insight into state-of-the-art morphological learning models with ever-improving prediction capabilities. Inasmuch as humans are the gold-standard in language learning and language use, one possible reason for current progress is that models are making predictions for more human-like reasons. The results here show that that intuition does not necessarily hold. The systems evaluated in this shared task were on the whole successful in their predictions but did not behave in a especially human-like manner.

Acknowledgements

We would like to thank shared task organizer Ekatrina Vylomova for her advice on logistics as well as Ellen Broselow, Jeff Heinz, and Charles Yang for their comments on this overview paper. The neural baseline system was trained on the Stony Brook SeaWulf HPC cluster, maintained by Research Computing and Cyberinfrastructure, and the Institute for Advanced Computational Science at Stony Brook University and made possible by NSF grant #1531492.

References

Shaima Alqattan. 2015. *Early phonological acquisition by Kuwaiti Arabic children*. Ph.D. thesis, Newcastle University.

Sharon L. Armstrong, Lila R. Gleitman, and Henry Gleitman. 1983. What some concepts might not be. *Cognition*, 13(3):263–308.

Mark Aronoff. 1976. *Word formation in generative grammar*. MIT Press, Cambridge, MA.

Harald Baayen. 1993. On frequency, transparency and productivity. In *Yearbook of morphology 1992*, pages 181–208. Springer, Dordrecht.

R Harald Baayen, Richard Piepenbrock, and H Van Rijn. 1993. The celex lexical database (cd-rom). linguistic data consortium. *Philadelphia, PA: University of Pennsylvania*.

Heike Behrens. 2006. The input–output relationship in first language acquisition. *Language and cognitive processes*, 21(1-3):2–24.

Caleb A Belth, Sarah RB Payne, Deniz Beser, Jordan Kodner, and Charles Yang. 2021. The greedy and recursive search for morphological productivity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Deniz Beser. 2021. Falling through the gaps: Neural architectures as models of morphological rule learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Michael Brent and Jay Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(1):31–44.

Roger Brown. 1973. *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Roger Brown, Courtney B. Cazden, and Ursula Bellugi. 1973. The child's grammar from I to III. In Charles A. Ferguson and Daniel I. Slobin, editors, *Studies of child language development*, pages 295–333. Holt, Rinehart and Winston, New York.

Joan L Bybee. 1985. *Morphology: A study of the relation between meaning and form*. John Benjamins, Amsterdam.

Harald Clahsen. 1990. Constraints on parameter setting: A grammatical analysis of some acquisition in stages in German child language. *Language Acquisition*, 1(4):361–391.

Harald Clahsen and Monika Rothweiler. 1993. Inflectional rules in children's grammars: Evidence from German participles. In *Yearbook of morphology 1992*, pages 1–34. Springer.

Harald Clahsen, Monika Rothweiler, Andreas Woest, and Gary Marcus. 1992. Regular and irregular inflection in the acquisition of German noun plurals. *Cognition*, 45:225–255.

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 Conference of the Association for Computational Linguistics, ACL*, pages 3868–3877.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. 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, Germany. Association for Computational Linguistics.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to german plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108.

Lisa Garnand Dawdy-Hesterberg and Janet Breckenridge Pierrehumbert. 2014. Learnability and generalisation of arabic broken plural nouns. *Language, cognition and neuroscience*, 29(10):1268–1282.

Bruce L Derwing and William J Baker. 1977. The psychological basis for morphological rules. In John Macnamara, editor, *Language learning and thought*, pages 85–110. Academic Press, New York.

Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Jeffrey Elman. 1998. Generalization, simple recurrent networks, and the emergence of structure. In *Proceedings of the twentieth annual conference of the Cognitive Science Society*, pages 543–548, Mahwah, NJ. Lawrence Erlbaum.

Hilke Elsen. 2002. The acquisition of German plurals. In *Morphology 2000: Selected Papers from the 9th Morphology Meeting, Vienna, 25-27 February 2000*, volume 218, page 117. John Benjamins Publishing.

Micha Elsner and Sara K. Court. 2022. OSU at SIGMORPHON 2022: Analogical Inflection With Rule Features. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.

M. Gareth Gaskell and William D. Marslen-Wilson. 2001. Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language*, 44(?):325–349.

Judith C Goodman, Philip S Dale, and Ping Li. 2008. Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language*, 35(3):515–531.

Kyle Gorman, Arya D McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but ok: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151.

Olivia Guest and Andrea E Martin. 2021. On logical inference over brains, behaviour, and artificial neural networks. *PsyArXiv preprint 10.31234/osf.io/tbmcg*.

Youssef A. Haddad. 2008. Pseudometathesis in Three Standard Arabic Broken-Plural Templates. *Word Structure*, 1:135–155.

Sophie Kern, Barbara Davis, and Inge Zink. 2009. From babbling to first words in four languages: Common trends across languages and individual differences.

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.

Jordan Kodner. 2019. Estimating child linguistic experience from historical corpora. *Glossa*, 4(1):122.

Jordan Kodner. 2022. Computational models of morphological learning. In *Oxford Research Encyclopedia of Linguistics*.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkuş, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanelov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Silvia Guriel-Agiashvili, Ritvan Karahodja, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Sheifer, Alexandra Serova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON-UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.

Henry Kučera and W Nelson Francis. 1967. *Computational analysis of present-day American-English*. Brown University Press, Providence, RI.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467, Cairo.

Brian MacWhinney. 1978. The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, pages 1–123.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Michael Maratsos. 2000. More overregularizations after all: New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen & Xu. *Journal of Child Language*, 27(1):183–212.

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 F Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development*.

Robert JC Maslen, Anna L Theakston, Elena VM Lieven, and Michael Tomasello. 2004. A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research*, 47(1319-1333).

Laia Mayol. 2007. Acquisition of irregular patterns in Spanish verbal morphology. In *Proceedings of the twelfth ESSLLI Student Session*, pages 1–11, Dublin.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina 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, Italy. Association for Computational Linguistics.

John J. McCarthy and Alan S. Prince. 1990. Foot and Word in Prosodic Morphology: The Arabic Broken Plural. *Natural Language & Linguistic Theory*, 8:209–283.

James L. McClelland and Karalyn Patterson. 2002. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6(11):465–472.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1745–1756.

Allen Parducci and Linda F. Perrett. 1971. Category rating scales: Effects of relative frequency of stimulus values. *Journal of Experimental Psychology*, 89(2):427–452.

Janet B Pierrehumbert. 2003. On frequency, transparency and productivity. In *Probabilistic Linguistics*, pages 177–228. MIT Press, Cambridge, MA.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. 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.

Steven Pinker and Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Science*, 6(11):456–463.

Akhilesh Kakolu Ramarao, Yulia Zinova, Kevin Tang, and Ruben van de Vijver. 2022. HeiMorph at SIGMORPHON 2022 Shared Task on Morphological Acquisition Trajectories. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.

Dorit Ravid and Rola Farah. 1999. Learning about noun plurals in early palestinian arabic. *First Language*, 19(56):187–206.

David E Rumelhart and James L McClelland. 1986. On learning the past tenses of english verbs.

Heba Salama and Sameh Alansary. 2017. Lexical growth in egyptian arabic speaking children: A corpus based study. *The Egyptian Journal of Language Engineering*, 4(1):29–34.

Carson T. Schütze. 2005. Thinking about what we are asking speakers to do. In Stephan Kepser and Marga Reis, editors, *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 457–485. Mouton de Gruyter, Berlin.

Simon Clematide Silvan Wehrli and Peter Makarov. 2022. CLUZH at SIGMORPHON 2022 Shared Tasks on Morpheme Segmentation and Inflection Generation. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.

Ingrid Sonnenstuhl and Axel Huth. 2002. Processing and representation of german-n plurals: A dual mechanism approach. *Brain and Language*, 81(1-3):276–290.

Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash. 2018. An Arabic morphological analyzer and generator with copious features. In *Proceedings of SIGMORPHON*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Richard Wiese. 1996. *The phonology of German*. Clarendon, Oxford.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Fei Xu and Steven Pinker. 1995. Weird past tense forms. *Journal of Child Language*, 22(3):531–556.

Charles Yang. 2002. *Knowledge and learning in natural language*. Oxford University Press on Demand.

Charles Yang. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

Charles Yang. 2020. Saussurean rhapsody: Systematicity and arbitrariness in language. In *The Oxford Handbook of the Lexicon*. Oxford University Press, USA.

Eugen Zaretsky and Benjamin P Lange. 2015. No matter how hard we try: Still no default plural marker in nonce nouns in modern high german. In *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, pages 153–178.

A Additional Analysis

The tables in this appendix present additional analyses referenced in the paper.

CLUZH	Match	SorW	SC+*ed*	Irreg	?
100	99.1	0.9	0.0	0.9	0.0
200	99.28	0.72	0.0	0.72	0.0
300	99.82	0.18	0.0	0.18	0.0
400	99.46	0.54	0.0	0.54	0.0
500	97.49	2.51	0.0	2.51	0.0
600	96.41	3.59	0.0	3.59	0.0
700	96.77	3.05	0.0	3.23	0.0
800	97.67	2.33	0.0	2.33	0.0
900	99.28	0.72	0.0	0.72	0.0
1000	97.49	2.51	0.0	2.51	0.0

HeiM	Match	SorW	SC+*ed*	Irreg	?
100	63.91	12.93	0.72	14.9	21.18
200	80.97	9.69	0.18	14.18	4.85
300	79.17	10.77	0.54	14.72	6.1
400	63.2	3.77	0.36	5.75	31.06
500	80.43	15.44	0.72	16.88	2.69
600	82.05	13.46	0.9	15.26	2.69
700	81.87	13.82	0.36	14.54	3.59
800	81.87	10.77	0.36	11.13	7.0
900	80.79	10.77	0.0	11.31	7.9
1000	88.15	5.39	0.18	6.46	5.39

OSU	Match	SorW	SC+*ed*	Irreg	?
100	79.17	8.98	3.77	15.44	5.39
200	87.97	4.13	1.8	7.18	4.85
300	91.74	3.41	0.9	5.21	3.05
400	92.82	2.33	0.18	3.23	3.95
500	90.66	3.05	0.9	4.85	4.49
600	92.82	3.77	0.36	4.31	2.87
700	93.36	3.05	0.36	3.77	2.87
800	94.61	3.59	0.0	3.59	1.8
900	97.49	1.8	0.0	1.8	0.72
1000	97.49	1.26	0.36	1.62	0.9

Table 15: Error type analysis for English regular verbs. *Match* = % correct or orthographic. *SorW* = % well-formed strong or weak irregular. *SC+ed* = % *-ed* is present but with a vowel change. *Irreg* = % all plausibly irregular patterns. *?* = nonsense output

CLUZH	Match	Other	Reg	*-ed*	?
100	4.65	4.65	88.37	88.37	2.33
200	2.33	4.65	93.02	93.02	0.0
300	2.33	4.65	93.02	93.02	0.0
400	2.33	2.33	95.35	95.35	0.0
500	9.3	6.98	83.72	83.72	0.0
600	13.95	4.65	81.4	81.4	0.0
700	6.98	4.65	83.72	86.05	2.33
800	9.3	4.65	86.05	86.05	0.0
900	4.65	2.33	93.02	93.02	0.0
1000	9.3	6.98	83.72	83.72	0.0

HeiM	Match	Other	Reg	*-ed*	?
100	9.3	18.6	58.14	69.77	2.33
200	11.63	9.3	69.77	74.42	4.65
300	13.95	18.6	55.81	62.79	4.65
400	9.3	9.3	60.47	81.4	0.0
500	6.98	37.21	46.51	51.16	4.65
600	11.63	39.53	32.56	41.86	6.98
700	9.3	30.23	51.16	58.14	2.33
800	4.65	20.93	60.47	72.09	2.33
900	6.98	16.28	60.47	74.42	2.33
1000	2.33	9.3	76.74	81.4	6.98

OSU	Match	Other	Reg	*-ed*	?
100	9.3	27.91	53.49	55.81	6.98
200	9.3	11.63	69.77	79.07	0.0
300	11.63	20.93	62.79	67.44	0.0
400	4.65	11.63	72.09	81.4	2.33
500	11.63	9.3	67.44	74.42	4.65
600	9.3	13.95	65.12	76.74	0.0
700	6.98	9.3	74.42	79.07	4.65
800	4.65	16.28	72.09	76.74	2.33
900	4.65	6.98	83.72	88.37	0.0
1000	2.33	4.65	88.37	90.7	2.33

Table 16: Error type analysis for English irregular verbs. *Match* = % correct. *Other* = % other plausible strong and weak irregulars. *Reg* = % "correct" regularized. *-ed* = % forms ending in -ed. *?* = other nonsense output

CLUZH	#	*-ed*	→*a*	→*u*	NC	Other	?
100	8	8	0	0	0	0	0
200	8	8	0	0	0	0	0
300	8	8	0	0	0	0	0
400	8	8	0	0	0	0	0
500	8	6	1	1	0	0	0
600	8	6	1	1	0	0	0
700	8	7	0	1	0	0	0
800	8	7	0	1	0	0	0
900	8	7	0	1	0	0	0
1000	8	4	1	3	0	0	0

HeiM	#	*-ed*	→*a*	→*u*	NC	Other	?
100	8	8	0	0	0	0	0
200	8	8	0	0	0	0	0
300	8	6	0	0	1	0	1
400	8	7	0	0	0	0	1
500	8	4	0	0	4	0	0
600	8	5	0	0	3	0	0
700	8	4	0	0	4	0	0
800	8	8	0	0	0	0	0
900	8	8	0	0	0	0	0
1000	8	8	0	0	0	0	0

OSU	#	*-ed*	→*a*	→*u*	NC	Other	?
100	8	8	0	0	0	0	0
200	8	8	0	0	0	0	0
300	8	4	2	2	0	0	0
400	8	3	1	2	0	2	0
500	8	7	0	1	0	0	0
600	8	5	1	1	0	1	0
700	8	7	0	1	0	0	0
800	8	7	0	1	0	0	0
900	8	7	0	1	0	0	0
1000	8	8	0	0	0	0	0

Table 17: Inflection type for English monosyllabic -*ing* verbs. *-ed* = regular. →*a* = *sing-sang*-type. →*u* = *sting-stung*-type. NC = no change. Other = other strong inflection. ? = nonsense inflection.

CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	1	0	0	0	0	1
P -(*e*)*n*	6	**191**	0	0	2	199
P -*er*	0	0	0	0	0	0
P -∅	0	0	0	0	0	0
P -*s*	0	0	0	0	**1**	1
P ?	0	0	0	0	0	0
Sum	7	191	0	0	3	201

HeiM	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	0	0	0	0	0	0
P -(*e*)*n*	7	190	0	0	3	200
P -*er*	0	0	0	0	0	0
P -∅	0	0	0	0	0	0
P -*s*	0	0	0	0	0	0
P ?	0	1	0	0	0	1
Sum	7	191	0	0	3	201

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	**2**	5	0	0	1	8
P -(*e*)*n*	3	181	0	0	2	186
P -*er*	1	0	0	0	0	1
P -∅	0	0	0	0	0	0
P -*s*	0	0	0	0	0	0
P ?	1	5	0	0	0	6
Sum	7	191	0	0	3	201

Table 18: German inflection confusion matrices at 600 training for FEM nouns only, disregarding Umlaut. *G* = Gold, *P* = Prediction.

CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	**134**	14	1	0	9	158
P -(*e*)*n*	0	**7**	0	0	0	7
P -*er*	0	0	0	0	0	0
P -∅	4	5	0	**102**	0	111
P -*s*	1	0	0	0	**7**	8
P ?	0	0	0	0	0	0
Sum	139	26	1	102	16	284

HeiM	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	126	12	1	1	11	151
P -(*e*)*n*	7	4	0	0	1	12
P -*er*	1	0	0	0	0	1
P -∅	4	10	0	99	1	114
P -*s*	1	0	0	0	2	3
P ?	0	0	0	2	1	3
Sum	139	26	1	102	16	284

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	125	13	1	0	9	148
P -(*e*)*n*	4	3	0	0	0	7
P -*er*	0	0	0	0	0	0
P -∅	5	10	0	99	0	114
P -*s*	1	0	0	0	6	7
P ?	4	0	0	3	1	8
Sum	139	26	1	102	16	284

Table 19: German inflection confusion matrices at 600 training for MASC nouns only, disregarding Umlaut. *G* = Gold, *P* = Prediction.

CLUZH	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	**33**	2	12	0	9	56
P -(*e*)*n*	0	0	0	1	0	1
P -*er*	0	0	3	0	0	3
P -∅	4	0	0	**46**	0	50
P -*s*	0	1	1	0	**3**	5
P ?	0	0	0	0	0	0
Sum	37	3	16	47	12	115

HeiM	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	28	1	11	3	5	48
P -(*e*)*n*	0	0	0	0	0	0
P -*er*	3	0	**4**	1	4	12
P -∅	5	0	0	43	0	48
P -*s*	0	1	1	0	1	3
P ?	1	1	0	0	2	4
Sum	37	3	16	47	12	115

OSU	G -*e*	G -(*e*)*n*	G -*er*	G -∅	G -*s*	Sum
P -*e*	28	1	12	1	8	50
P -(*e*)*n*	0	0	0	0	0	0
P -*er*	1	0	3	1	0	5
P -∅	6	0	1	43	1	51
P -*s*	1	1	0	1	2	5
P ?	1	1	0	1	1	4
Sum	37	3	16	47	12	115

Table 20: German inflection confusion matrices at 600 training for NEUT nouns only, disregarding Umlaut. *G* = Gold, *P* = Prediction.

	S→S	S→B	B→S	B→B
CLUZH F	7	29	45	48
HeiM F	1	9	21	3
OSU F	2	13	23	0
CLUZH M	0	13	23	4
HeiM M	9	14	66	62
OSU M	11	18	41	57
CLUZH HUM	0	3	16	14
HeiM HUM	0	4	15	15
OSU HUM	2	1	15	16
CLUZH NHUM	7	39	52	38
HeiM NHUM	10	19	72	50
OSU NHUM	11	30	49	41

Table 21: Arabic error types at 1000 training by gender and rationality.

CLUZH	Gold SF	Gold SM	Gold B	Sum
Pred SF	139	0	23	162
Pred SM	0	0	0	0
Pred B	13	0	49	62
Pred ?	4	0	1	5
Sum	156	0	73	229

HeiM	Gold SF	Gold SM	Gold B	Sum
Pred SF	**140**	0	21	161
Pred SM	1	0	0	1
Pred B	9	0	**51**	60
Pred ?	6	0	1	7
Sum	156	0	73	229

OSU	Gold SF	Gold SM	Gold B	Sum
Pred SF	138	0	23	161
Pred SM	2	0	0	2
Pred B	13	0	45	58
Pred ?	3	0	5	8
Sum	156	0	73	229

Table 22: Arabic inflection confusion matrices for each submitted system at 1000 training. FEM nouns only.

CLUZH	Gold SF	Gold SM	Gold B	Sum
Pred SF	74	5	29	108
Pred SM	2	**51**	16	69
Pred B	25	4	**157**	186
Pred ?	0	2	6	8
Sum	101	62	208	371

HeiM	Gold SF	Gold SM	Gold B	Sum
Pred SF	**87**	7	51	145
Pred SM	2	43	15	60
Pred B	9	5	126	140
Pred ?	3	7	16	26
Sum	101	62	208	371

OSU	Gold SF	Gold SM	Gold B	Sum
Pred SF	80	8	26	114
Pred SM	3	50	15	68
Pred B	16	2	**157**	175
Pred ?	2	2	10	14
Sum	101	62	208	371

Table 23: Arabic inflection confusion matrices for each submitted system at 1000 training. MASC nouns only.

CLUZH	Gold SF	Gold SM	Gold B	Sum
Pred SF	**14**	0	1	15
Pred SM	0	**48**	15	63
Pred B	1	2	**24**	27
Pred ?	0	0	3	3
Sum	15	50	43	108

HeiM	Gold SF	Gold SM	Gold B	Sum
Pred SF	12	0	0	12
Pred SM	0	43	15	58
Pred B	2	2	16	20
Pred ?	1	5	12	18
Sum	15	50	43	108

OSU	Gold SF	Gold SM	Gold B	Sum
Pred SF	13	2	0	15
Pred SM	0	47	15	62
Pred B	1	0	**24**	25
Pred ?	1	1	4	6
Sum	15	50	43	108

Table 24: Arabic inflection confusion matrices for each submitted system at 1000 training. HUM nouns only.

CLUZH	Gold SF	Gold SM	Gold B	Sum
Pred SF	199	5	51	255
Pred SM	2	**3**	1	6
Pred B	37	2	182	**221**
Pred ?	4	2	4	10
Sum	242	12	238	492

HeiM	Gold SF	Gold SM	Gold B	Sum
Pred SF	**215**	7	72	294
Pred SM	3	0	0	3
Pred B	16	3	161	180
Pred ?	8	2	5	15
Sum	242	12	238	492

OSU	Gold SF	Gold SM	Gold B	Sum
Pred SF	205	6	49	260
Pred SM	5	**3**	0	8
Pred B	28	2	178	208
Pred ?	4	1	11	16
Sum	242	12	238	492

Table 25: Arabic inflection confusion matrices for each submitted system at 1000 training. NHUM nouns only.