The SIGMORPHON 2016 Shared Task—Morphological Reinflection

Ryan Cotterell
Dept. of Computer Science
Johns Hopkins University
ryan.cotterell@jhu.edu

Christo Kirov
Dept. of Computer Science
Johns Hopkins University
ckirov@gmail.com

John Sylak-Glassman
Dept. of Computer Science
Johns Hopkins University
jcsg@jhu.edu

David Yarowsky
Dept. of Computer Science
Johns Hopkins University
yarowsky@jhu.edu

Jason Eisner
Dept. of Computer Science
Johns Hopkins University
jason@jhu.edu

Mans Hulden
Dept. of Linguistics
University of Colorado
mans.hulden@colorado.edu

Abstract

The 2016 SIGMORPHON Shared Task was devoted to the problem of morphological reinflection. It introduced morphological datasets for 10 languages with diverse typological characteristics. The shared task drew submissions from 9 teams representing 11 institutions reflecting a variety of approaches to addressing supervised learning of reinflection. For the simplest task, inflection generation from lemmas, the best system averaged 95.56% exact-match accuracy across all languages, ranging from Maltese (88.99%) to Hungarian (99.30%). With the relatively large training datasets provided, recurrent neural network architectures consistently performed best—in fact, there was a significant margin between neural and non-neural approaches. The best neural approach, averaged over all tasks and languages, outperformed the best non-neural one by 13.76% absolute; on individual tasks and languages the gap in accuracy sometimes exceeded 60%. Overall, the results show a strong state of the art, and serve as encouragement for future shared tasks that explore morphological analysis and generation with varying degrees of supervision.

1 Introduction

Many languages use systems of rich overt morphological marking in the form of affixes (i.e. suffixes, prefixes, and infixes) to convey syntactic and semantic distinctions. For example, each English count noun has both singular and plural forms (e.g. *robot*/*robots*, *process*/*processes*), and these are known as the inflected forms of the noun. While English has relatively little inflectional morphology, Russian nouns, for example, can have a total of 10 distinct word forms for any given lemma and 30 for an imperfective verb.[1] In the extreme, Kibrik (1998) demonstrates that even by a conservative count, a verb conjugation in Archi (Nakh-Daghestanian) consists of 1,725 forms, and if all sources of complexity are considered, a single verb lemma may give rise to up to 1,502,839 distinct forms. The fact that inflected forms are systematically related to each other, as shown in Figure 1, is what allows humans to generate and analyze words despite this level of morphological complexity.

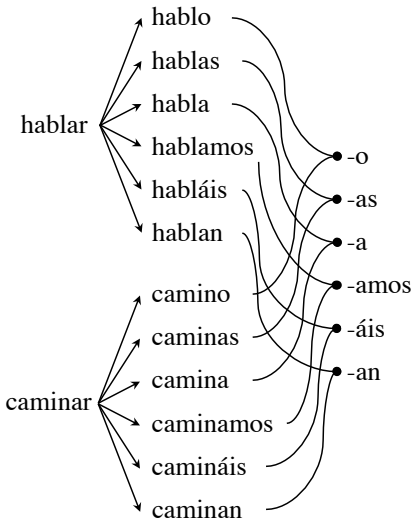

Figure 1: The relatedness of inflected forms, such as the present indicative paradigm of the Spanish verbs *hablar* 'speak' and *caminar* 'walk,' allows generalizations about the shape and affixal content of the paradigm to be extracted.

A core problem that arises in languages with rich morphology is data sparsity. When a single lexical item can appear in many different word

[1] This latter figure rises to 52 if the entire imperfective-perfective pair (e.g. *govorit'/skazat'* 'speak, tell') is considered to be a single lemma.

Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 10–22,
Berlin, Germany, August 11, 2016.

forms, the probability of encountering any single word form decreases, reducing the effectiveness of frequency-based techniques in performing tasks like word alignment and language modeling (Koehn, 2010; Duh and Kirchhoff, 2004). Techniques like lemmatization and stemming can ameliorate data sparsity (Goldwater and McClosky, 2005), but these rely on morphological knowledge, particularly the mapping from inflected forms to lemmas and the list of morphs together with their ordering. Developing systems that can accurately learn and capture these mappings, overt affixes, and the principles that govern how those affixes combine is crucial to maximizing the cross-linguistic capabilities of most human language technology.

The goal of the 2016 SIGMORPHON Shared Task[2] was to spur the development of systems that could accurately generate morphologically inflected words for a set of 10 languages based on a range of training parameters. These 10 languages included low resource languages with diverse morphological characteristics, and the training parameters reflected a significant expansion upon the traditional task of predicting a full paradigm from a lemma. Of the systems submitted, the neural network-based systems performed best, clearly demonstrating the effectiveness of recurrent neural networks (RNNs) for morphological generation and analysis.

We are releasing the shared task data and evaluation scripts for use in future research.

2 Tasks, Tracks, and Evaluation

Up to the present, the task of morphological inflection has been narrowly defined as the generation of a complete inflectional paradigm from a lemma, based on training from a corpus of complete paradigms.[3] This task implicitly assumes the availability of a traditional dictionary or gazetteer, does not require explicit morphological analysis, and, though it mimics a common task in second language (L2) pedagogy, it is not a realistic learning setting for first language (L1) acquisition.

Systems developed for the 2016 Shared Task had to carry out *reinflection* of an already inflected form. This involved *analysis* of an already inflected word form, together with *synthesis* of a different inflection of that form. The systems had to learn from limited data: they were not given complete paradigms to train on, nor a dictionary of lemmas.

	Task 1	**Task 2**	**Task 3**
Lemma	run	—	—
Source tag	—	PAST	—
Source form	—	ran	ran
Target tag	PRESPART	PRESPART	PRESPART
Target form	running	running	running
Lemma	decir	—	—
Source tag	—	PRESENT1S	—
Source form	—	digo	digo
Target tag	FUTURE2S	FUTURE2S	FUTURE2S
Target form	dirás	dirás	dirás

Table 1: Systems were required to generate the target form, given the information above the line. Two examples are shown for each task—one in English and one in Spanish. Task 1 is inflection; tasks 2–3 are reinflection.

	Restricted	Standard	Bonus
Task 1	1	1	1, M
Task 2	2	1, 2	1, 2, M
Task 3	3	1, 2, 3	1, 2, 3, M

Table 2: Datasets that were permitted for each task under each condition. Numbers indicate a dataset from that respective task, e.g. '1' is the dataset from Task 1, and 'M' indicates bonus monolingual text from Wikipedia dumps.

Specifically, systems competed on the three tasks illustrated in Table 1, of increasing difficulty. Notice that each task can be regarded as mapping a source string to a target string, with other input arguments (such as the target tag) that specify which version of the mapping is desired.

For each language and each task, participants were provided with supervised training data: a collection of input tuples, each paired with the correct output string (target form).

Each system could compete on a task under any of three tracks (Table 2). Under the restricted track, only data for that task could be used, while for the standard track, data from that task and any from a lower task could be used. The bonus track was the same as the standard track, but allowed the use of monolingual data in the form of Wikipedia dumps from 2 November 2015.[4]

Each system was required to produce, for every input given at test time, either a single string or a ranked list of up to 20 predicted strings for each task. Systems were compared on the follow-

[2] Official website: http://ryancotterell.github.io/sigmorphon2016/

[3] A paradigm is defined here as the set of inflected word forms associated with a single lemma (or lexeme), for example, a noun declension or verb conjugation.

[4] https://dumps.wikimedia.org/backup-index.html

ing metrics, averaged over all inputs:

- Accuracy: 1 if the top predicted string was correct, else 0
- Levenshtein distance: Unweighted edit distance between the top predicted string and the correct form
- Reciprocal rank: $1/(1 + \text{rank}_i)$, where rank_i is the rank of the correct string, or 0 if the correct string is not on the list

The third metric allows a system to get partial credit for including a correct answer on its list, preferably at or near the top.

3 Data

3.1 Languages and Typological Characteristics

Datasets from 10 languages were used. Of these, 2 were held as surprise languages whose identity and data were only released at evaluation time.

- *Standard Release*: Arabic, Finnish, Georgian, German, Navajo, Russian, Spanish, and Turkish
- *Surprise*: Hungarian and Maltese

Finnish, German, and Spanish have been the subject of much recent work, due to data made available by Durrett and DeNero (2013), while the other datasets used in the shared task are released here for the first time. For all languages, the word forms in the data are orthographic (not phonological) strings in the native script, except in the case of Arabic, where we used the romanized forms available from Wiktionary. An accented letter is treated as a single character. Descriptive statistics of the data are provided in Table 3.

The typological character of these languages varies widely. German and Spanish inflection generation has been studied extensively, and the morphological character of the languages is similar: Both are suffixing and involve internal stem changes (e.g., `a` ↦ `ä`, `e` ↦ `ie`, respectively). Russian can be added to this group, but with consonantal rather than vocalic stem alternations. Finnish, Hungarian, and Turkish are all agglutinating, almost exclusively suffixing, and have vowel harmony systems. Georgian exhibits complex patterns of verbal agreement for which it utilizes circumfixal morphology, i.e. simultaneous prefixation and suffixation (Aronson, 1990).

	Split	Pairs	Lem	Full	T2T	I-Tag	O-Tag	Sync
Ar	train	12616	2130	225	1.57	72.23	56.57	1.10
	dev	1596	1081	220	1.08	9.13	7.26	1.08
	test	15643	2150	230	1.71	87.06	69.22	1.12
Fi	train	12764	9855	95	5.70	142.15	134.36	1.01
	dev	1599	1546	91	1.51	19.28	18.60	1.01
	test	23878	15128	95	9.87	261.00	251.34	1.01
Ge	train	12390	4246	90	14.02	152.38	137.67	1.06
	dev	1591	1274	77	5.31	24.25	23.40	1.03
	test	21813	4622	90	15.32	279.43	242.36	1.08
Ge	train	12689	6703	99	7.76	246.19	129.48	1.44
	dev	1599	1470	98	1.80	30.82	16.32	1.37
	test	15777	7277	100	9.48	300.49	159.37	1.50
Hu	train	18206	1508	87	9.05	231.13	211.70	1.04
	dev	2381	1196	83	2.14	30.27	29.04	1.02
	test	2360	1186	84	2.09	29.52	28.78	1.02
Ma	train	19125	1453	3607	1.00	6.00	6.01	1.00
	dev	2398	1033	1900	1.00	1.62	1.61	1.00
	test	2399	1055	1928	1.00	1.61	1.59	1.00
Na	train	10478	355	54	17.48	310.55	194.03	1.47
	dev	1550	326	47	2.80	44.93	33.69	1.17
	test	686	233	42	2.89	25.56	16.33	1.12
Ru	train	12663	7941	83	10.32	182.25	152.56	1.07
	dev	1597	1492	78	2.36	23.69	20.74	1.06
	test	23445	10560	86	17.87	320.28	282.47	1.09
Sp	train	12725	5872	84	3.24	186.38	151.48	1.06
	dev	1599	1406	83	1.41	23.08	19.26	1.07
	test	23743	7850	84	5.42	342.72	286.06	1.06
Tu	train	12645	2353	190	1.81	79.82	67.62	1.08
	dev	1599	1125	170	1.09	11.15	9.57	1.06
	test	1598	1128	170	1.08	10.99	9.57	1.05

Table 3: Descriptive statistics on data released to shared task participants. Figures represent averages across tasks. Abbreviations in the headers: 'Lem' = lemmas, 'Full' = number of full tags, T2T = average occurrences of tag-to-tag pairs, I-Tag & O-Tag = average occurrences of each input or output tag, resp., and 'Sync' = average forms per tag (syncretism).

Navajo, like other Athabaskan languages, has primarily prefixing verbal morphology with consonant harmony among its sibilants (Rice, 2000; Hansson, 2010). Arabic and Maltese, both Semitic languages, utilize templatic, non-concatenative morphology. Maltese, due partly to its contact with Italian, also uses concatenative morphology (Camilleri, 2013).

3.2 Quantifying Morphological Processes

It is helpful to understand how often each language makes use of different morphological processes and where they apply. In lieu of a more careful analysis, here we use a simple heuristic to estimate how often inflection involves prefix changes, stem-internal changes (apophony), or suffix changes (Table 4). We assume that each word form in the training data can be divided into three parts—prefix, stem and suffix—with the prefix and suffix possibly being empty.

To align a source form with a target form, we pad both of them with – symbols at their start and/or end (but never in the middle) so that they have equal length. As there are multiple ways

Language	Prefix	Stem	Suffix
Arabic	68.52	37.04	88.24
Finnish	0.02	12.33	96.16
Georgian	4.46	0.41	92.47
German	0.84	3.32	89.19
Hungarian	0.00	0.08	99.79
Maltese	48.81	11.05	98.74
Navajo	77.64	18.38	26.40
Russian	0.66	7.70	85.00
Spanish	0.09	3.25	90.74
Turkish	0.21	1.12	98.74

Table 4: Percentage of inflected word forms that have modified each part of the lemma, as estimated from the "lemma ↦ inflected" pairs in task 1 training data. A sum $< 100\%$ for a language implies that sometimes source and target forms are identical; a sum $> 100\%$ implies that sometimes multiple parts are modified.

to pad, we choose the alignment that results in minimum Hamming distance between these equal-length padded strings, i.e., characters at corresponding positions should disagree as rarely as possible. For example, we align the German verb forms `brennen` 'burn' and `gebrannt` 'burnt' as follows:

```
--brennen
gebrannt-
```

From this aligned string pair, we heuristically split off a prefix pair before the first matching character (∅ ↦ `ge`), and a suffix pair after the last matching character (`en` ↦ `t`). What is left is presumed to be the stem pair (`brenn` ↦ `brann`):

Pref.	Stem	Suff.
	`brenn`	`en`
`ge`	`brann`	`t`

We conclude that when correctly mapping this source form to this target form, the prefix, stem, and suffix parts all change. In what fraction of training examples does each change, according to this heuristic? Statistics for each language (based on task 1) are shown in Table 4.

The figures roughly coincide with our expectations. Finnish, Hungarian, Russian, Spanish, and Turkish are largely or exclusively suffixing. The tiny positive number for Finnish prefixation is due to a single erroneous pair in the dataset. The large rate of stem-changing in Finnish is due to the phenomenon of consonant gradation, where stems undergo specific consonant changes in certain inflected forms. Navajo is primarily prefixing,[5] and Arabic exhibits a large number of "stem-internal" changes due to its templatic morphology. Maltese, while also templatic, shows fewer stem-changing operations than Arabic overall, likely a result of influence from non-Semitic languages. Georgian circumfixal processes are reflected in an above-average number of prefixes. German has some prefixing, where essentially the only formation that counts as such is the circumfix `ge_____t` for forming the past participle.

3.3 Data Sources and Annotation Scheme

Most data used in the shared task came from the English edition of Wiktionary.[6] Wiktionary is a crowdsourced, broadly multilingual dictionary with content from many languages (e.g. Spanish, Navajo, Georgian) presented within editions tailored to different reader populations (e.g. English-speaking, Spanish-speaking). Kirov et al. (2016) describe the process of extracting lemmas and inflected wordforms from Wiktionary, associating them with morphological labels from Wiktionary, and mapping those labels to a universalized annotation scheme for inflectional morphology called the UniMorph Schema (Sylak-Glassman et al., 2015b).

The goal of the UniMorph Schema is to encode the meaning captured by inflectional morphology across the world's languages, both high- and low-resource. The schema decomposes the morphological labels into universal attribute-value pairs. As an example, consider again Table 1. The FUT2S label for a Spanish future tense second-person singular verb form, such as `dirás`, is decomposed into [POS=VERB, mood=INDICATIVE, tense=FUTURE, person=2, number=SINGULAR].

The accuracy of data extraction and label association for data from Wiktionary was verified according to the process described in Kirov et al. (2016). However, verifying the full linguistic accuracy of the data was beyond the scope of preparation for the task, and errors that resulted from the original input of data by crowdsourced authors remained in some cases. These are noted in several of the system description papers. The full dataset from the English edition of Wiktionary, which in-

[5] The Navajo verb stem is always a single syllable appearing in final position, causing our heuristic to misclassify many stem changes as suffixal. In reality, verb suffixes are very rare in Navajo (Young and Morgan, 1987).

[6] https://en.wiktionary.org

cludes data from 350 languages, ≈977,000 lemmas, and ≈14.7 million inflected word forms, is available at `unimorph.org`, along with detailed documentation on the UniMorph Schema and links to the references cited above.

The Maltese data came from the Ġabra open lexicon[7] (Camilleri, 2013), and the descriptive features for inflected word forms were mapped to features in the UniMorph Schema similarly to data from Wiktionary. This data did not go through the verification process noted for the Wiktionary data.

Descriptive statistics for the data released to shared task participants are given in Table 3.

4 Previous Work

Much previous work on computational approaches to inflectional morphology has focused on a special case of reinflection, where the input form is always the lemma (i.e. the citation form). Thus, the task is to generate all inflections in a paradigm from the lemma and often goes by the name of *paradigm completion* in the literature. There has been a flurry of recent work in this vein: Durrett and DeNero (2013) heuristically extracted transformational rules and learned a statistical model to apply the rules, Nicolai et al. (2015) tackled the problem using standard tools from discriminative string transduction, Ahlberg et al. (2015) used a finite-state construction to extract complete candidate inflections at the paradigm level and then train a classifier, Faruqui et al. (2016) applied a neural sequence-to-sequence architecture (Sutskever et al., 2014) to the problem.

In contrast to paradigm completion, the task of reinflection is harder as it may require both morphologically analyzing the source form and transducing it to the target form. In addition, the training set may include only partial paradigms. However, many of the approaches taken by the shared task participants drew inspiration from work on paradigm completion.

Some work, however, has considered full reinflection. For example, Dreyer and Eisner (2009) and Cotterell et al. (2015) apply graphical models with string-valued variables to model the paradigm jointly. In such models it is possible to predict values for cells in the paradigm conditioned on sets of other cells, which are not required to include the lemma.

[7] `http://mlrs.research.um.edu.mt/resources/gabra/`

5 Baseline System

To support participants in the shared task, we provided a baseline system that solves all tasks in the standard track (see Tables 1–2).

Given the input string (source form), the system predicts a left-to-right sequence of edits that convert it to an output string—hopefully the correct target form. For example, one sequence of edits that could be legally applied to the Finnish input `katossa` is *copy*, *copy*, *copy*, *insert*(`t`), *copy*, *delete*(3). This results in the output `katto`, via the following alignment:

1	2	3	4	5	6
k	a	t	–	o	ssa
k	a	t	t	o	–

In general, each edit has the form *copy*, *insert*(string), *delete*(number), or *subst*(string), where *subst*(w) has the same effect as *delete*($|w|$) followed by *insert*(w).

The system treats edit sequence prediction as a sequential decision-making problem, greedily choosing each edit action given the previously chosen actions. This choice is made by a deterministic classifier that is trained to choose the correct edit on the assumption that that all previous edits on this input string were correctly chosen.

To prepare training data for the classifier, each supervised word pair in training data was aligned to produce a desired sequence of edits, such as the 6-edit sequence above, which corresponds to 6 supervised training examples. This was done by first producing a character-to-character alignment of the source and target forms (`katossa`, `katto`), using an iterative Markov Chain Monte Carlo method,and then combining consecutive deletions, insertions, or substitutions into a single compound edit. For example, *delete*(3) above was obtained by combining the consecutive deletions of `s`, `s`, and `a`.

The system uses a linear multi-class classifier that is trained using the averaged perceptron method (Freund and Schapire, 1999). The classifier considers the following binary features at each position:

- The previous 1, 2, and 3 input characters, e.g. `t`, `at`, `kat` for the 4th edit in the example.
- The previous 1, 2, and 3 output characters, e.g. `t`, `tt`, `att` for the 5th edit.

- The following 1, 2, and 3 input characters, e.g. `o`, `os`, `oss` for the 3rd edit.
- The previous edit. (The possible forms were given above.)
- The UniMorph morphosyntactic features of the source tag S or the target tag T (according to what type of mapping we are building—see below). For example, when lemmatizing `katossa` into `katto` as in the example above, $S = \left[\text{POS=NOUN, case=IN+ESS, number=SINGULAR}\right]$, yielding 3 morphosyntactic features.
- Each conjunction of two features from the above list where the first feature in the combination is a morphosyntactic feature and the second is not.

For task 1, we must edit from LEMMA $\rightarrow T$. We train a separate edit classifier for each part-of-speech, including the morphosyntactic description of T as features of the classifier. For task 2, we must map from $S \rightarrow T$. We do so by lemmatizing $S \rightarrow$ LEMMA (lemmatization) and then reinflecting LEMMA $\rightarrow T$ via the task 1 system.[8] For the lemmatization step, we again train a separate edit classifier for each part-of-speech, which now draws on source tag S features. For task 3, we build an additional classifier to analyze the source form to its morphosyntactic description S (using training data from all tasks, as allowed in the standard track). This classifier uses substrings of the word form as its features, and is also implemented by an averaged perceptron. The classifier treats each unique sequence of feature-value pairs as a separate class. Task 3 is then solved by first recovering the source tag and then applying the task 2 system.

The baseline system performs no tuning of parameters or feature selection. The averaged perceptron is not trained with early stopping or other regularization and simply runs for 10 iterations or until the data are separated. The results of the baseline system are given in Table 5. Most participants in the shared task were able to outperform the baseline, often by a significant margin.

[8]Note that at training time, we know the correct lemma for S thanks to the task 1 data, which is permitted for use by task 2 in the standard track. This is also why task 2 is permitted to use the trained task 1 system.

Language	Task 1	Task 2	Task 3
Arabic	66.96	55.00	45.15
Finnish	64.45	59.59	56.95
Georgian	89.12	86.66	85.12
German	89.44	87.62	80.13
Hungarian	73.42	72.78	71.70
Maltese	38.49	27.54	26.00
Navajo	53.06	47.59	44.96
Russian	88.65	84.68	79.55
Spanish	95.72	94.54	87.51
Turkish	59.60	57.63	55.25

Table 5: Accuracy results for the baseline system on the standard track test set.

6 System Descriptions

The shared task received a diverse set of submissions with a total of 11 systems from 9 teams representing 11 different institutions. For the sake of clarity, we have grouped the submissions into three camps.

The first camp adopted a pipelined approach similar to that of the baseline system provided. They first employed an unsupervised alignment algorithm on the source-target pairs in the training data to extract a set of edit operations. After extraction, they applied a discriminative model to apply the changes. The transduction models limited themselves to monotonic transduction and, thus, could be encoded through weighted finite-state machine (Mohri et al., 2002).

The second camp focused on neural approaches, building on the recent success of neural sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2014). Recently, Faruqui et al. (2016) found moderate success applying such networks to the inflection task (our task 1). The neural systems were the top performers.

Finally, the third camp relied on linguistically-inspired heuristic means to reduce the structured task of reinflection to a more reasonable multi-way classification task that could be handled with standard machine learning tools.

6.1 Camp 1: Align and Transduce

Most of the systems in this camp drew inspiration from the work of Durrett and DeNero (2013), who extracted a set of edit operations and applied the transformations with a semi-Markov

CRF (Sarawagi and Cohen, 2004).

EHU EHU (Alegria and Etxeberria, 2016) took an approach based on standard grapheme-to-phoneme machinery. They extend the Phonetisaurus (Novak et al., 2012) toolkit, based on the OpenFST WFST library (Allauzen et al., 2007), to the task of morphological reinflection. Their system is organized as a pipeline. Given pairs of input and output strings, the first step involves an unsupervised algorithm to extract an alignment (many-to-one or one-to-many). Then, they train the weights of the WFSTs using the imputed alignments, introducing morphological tags as symbols on the input side of the transduction.

Alberta The Alberta system (Nicolai et al., 2016) is derived from the earlier work by Nicolai et al. (2015) and is methodologically quite similar to that of EHU—an unsupervised alignment model is first applied to the training pairs to impute an alignment. In this case, they employ the M2M-aligner (Jiampojamarn et al., 2007). In contrast to EHU, Nicolai et al. (2016) do allow many-to-many alignments. After computing the alignments, they discriminatively learn a string-to-string mapping using the DirectTL+ model (Jiampojamarn et al., 2008). This model is state-of-the-art for the grapheme-to-phoneme task and is very similar to the EHU system in that it assumes a monotonic alignment and could therefore be encoded as a WFST. Despite the similarity to the EHU system, the model performs much better overall. This increase in performance may be attributable to the extensive use of language-specific heuristics, detailed in the paper, or the application of a discriminative reranker.

Colorado The Colorado system (Liu and Mao, 2016) took the same general tack as the previous two systems—they used a pipelined approach that first discovered an alignment between the string pairs and then discriminatively trained a transduction. The alignment algorithm employed is the same as that of the baseline system, which relies on a rich-get-richer scheme based on the Chinese restaurant process (Sudoh et al., 2013), as discussed in §5. After obtaining the alignments, they extracted edit operations based on the alignments and used a semi-Markov CRF to apply the edits in a manner very similar to the work of Durrett and DeNero (2013).

OSU The OSU system (King, 2016) also used a pipelined approach. They first extracted sequences of edit operations using Hirschberg's algorithm (Hirschberg, 1975). This reduces the string-to-string mapping problem to a sequence tagging problem. Like the Colorado system, they followed Durrett and DeNero (2013) and used a semi-Markov CRF to apply the edit operations. In contrast to Durrett and DeNero (2013), who employed a 0th-order model, the OSU system used a 1st-order model. A major drawback of the system was the cost of inference. The unpruned set of edit operations had over 500 elements. As the cost of inference in the model is quadratic in the size of the state space (the number of edit operations), this created a significant slowdown with over 15 days required to train in some cases.

6.2 Camp 2: Revenge of the RNN

A surprising result of the shared task is the large performance gap between the top performing neural models and the rest of the pack. Indeed, the results of Faruqui et al. (2016) on the task of morphological inflection only yielded modest gains in some languages. However, the best neural approach outperformed the best non-neural approach by an average (over languages) of 13.76% absolute accuracy, and at most by 60.04%!

LMU The LMU system (Kann and Schütze, 2016) was the all-around best performing system in the shared task. The system builds off of the encoder-decoder model for machine translation (Sutskever et al., 2014) with a soft attention mechanism (Bahdanau et al., 2014). The architecture is identical to the RNN encoder-decoder architecture of Bahdanau et al. (2014)—a stacked GRU (Cho et al., 2014). The key innovation is in the formatting of the data. The input word along with both the source and target tags were fed into the network as a *single string* and trained to predict the target string. In effect, this means that if there are n elements in the paradigm, there is a single model for all n^2 possible reinflectional mappings. Thus, the architecture shares parameters among all reinflections, using a single encoder and a single decoder.

BIU-MIT The BIU-MIT (Aharoni et al., 2016) team submitted two systems. Their first model, like LMU, built upon the sequence-to-sequence architecture (Sutskever et al., 2014; Bahdanau et al., 2014; Faruqui et al., 2016), but with several im-

System	Standard			Restricted			Bonus		
	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3	Task 1	Task 2	Task 3
LMU-1	1.0 (95.56)	1.0 (96.35)	1.0 (95.83)	1.0 (95.56)	1.0 (95.34)	1.0 (90.95)	1.0 (96.71)	1.0 (96.35)	1.0 (95.83)
LMU-2	2.0 (95.56)	2.0 (96.23)	2.0 (95.83)	2.0 (95.56)	2.0 (95.27)	2.0 (90.95)	—	—	—
BIU/MIT-1	—	—	—	4.2 (92.65)	5.2 (77.70)	3.8 (76.39)	—	—	—
BIU/MIT-2	—	—	—	4.2 (93.00)	4.2 (81.29)	—	—	—	—
HEL	—	—	—	3.9 (92.89)	3.5 (86.30)	3.2 (86.48)	—	—	—
MSU	3.8 (84.06)	3.6 (86.06)	3.8 (84.87)	6.2 (84.06)	6.0 (79.68)	6.2 (62.16)	—	—	—
CU	4.6 (81.02)	5.0 (72.98)	5.0 (71.75)	7.3 (81.02)	6.9 (69.89)	5.5 (67.91)	—	—	—
EHU	5.5 (79.24)	—	—	8.0 (79.67)	—	—	—	—	—
COL/NYU	6.5 (67.86)	4.7 (75.59)	4.8 (67.61)	9.2 (67.86)	7.2 (77.34)	6.3 (53.56)	2.7 (72.30)	2.5 (71.74)	2.6 (67.61)
OSU	—	—	—	9.0 (72.71)	—	—	—	—	—
UA	4.6 (81.83)	4.7 (74.06)	4.4 (71.23)	—	—	—	2.3 (79.95)	2.5 (71.56)	2.4 (70.04)
ORACLE.E	97.49	98.15	97.97	98.32	97.84	95.80	98.14	97.80	97.57

Table 6: Summary of results, showing average rank (with respect to other competitors) and average accuracy (equally weighted average over the 10 languages and marked in parentheses) by system. Oracle ensemble (ORACLE.E) accuracy represents the probability that at least one of the submitted systems predicted the correct form.

provements. Most importantly, they augment the encoder with a bidirectional LSTM to get a more informative representation of the context and they represent individual morphosyntactic attributes as well. In addition, they include template-inspired components to better cope with the templatic morphology of Arabic and Maltese. The second architecture, while also neural, more radically departs from previously proposed sequence-to-sequence models. The aligner from the baseline system is used to create a series of edit actions, similar to the systems in Camp 1. Rather than use a CRF, the BIU-MIT team predicted the sequence of edit actions using a neural model, much in the same way as a transition-based LSTM parser does (Dyer et al., 2015; Kiperwasser and Goldberg, 2016). The architectural consequence of this is that it replaces the soft alignment mechanism of (Bahdanau et al., 2014) with a hard attention mechanism, similar to Rastogi et al. (2016).

Helsinki The Helsinki system (Östling, 2016), like LMU and BIU-MIT, built off of the sequence-to-sequence architecture, augmenting the system with several innovations. First, a single decoder was used, rather than a unique one for all possible morphological tags, which allows for additional parameter sharing, similar to LMU. More LSTM layers were also added to the decoder, creating a deeper network. Finally, a convolutional layer over the character inputs was used, which was found to significantly increase performance over models without the convolutional layers.

6.3 Camp 3: Time for Some Linguistics

The third camp relied on linguistics-inspired heuristics to reduce the problem to multi-way classification. This camp is less unified than the other two, as both teams used very different heuristics.

Columbia – New York University Abu Dhabi The system developed jointly by Columbia and NYUAD (Taji et al., 2016) is based on the work of Eskander et al. (2013). It is unique among the submitted systems in that the first step in the pipeline is segmentation of the input words into prefixes, stems, and suffixes. Prefixes and suffixes are directly associated with morphological features. Stems within paradigms are further processed, using either linguistic intuitions or an empirical approach based on string alignments, to extract the stem letters that undergo changes across inflections. The extracted patterns are intended to capture stem-internal changes, such as vowel changes in Arabic. Reinflection is performed by selecting a set of changes to apply to a stem, and attaching appropriate affixes to the result.

Moscow State The Moscow State system (Sorokin, 2016) is derived from the work of Ahlberg et al. (2014) and Ahlberg et al. (2015). The general idea is to use finite-state techniques to compactly model all paradigms in an abstract form called an 'abstract paradigm'. Roughly speaking, an abstract paradigm is a set of rule transformations that derive all slots from the shared string subsequences present in each slot. Their method relies on the computation of longest common subsequence (Gusfield, 1997) to derive the abstract paradigms, which is similar to its use in the related task of lemmatization (Chrupała et al., 2008;

Müller et al., 2015). Once a complete set of abstract paradigms has been extracted from the data, the problem is reduced to multi-way classification, where the goal is to select which abstract paradigm should be applied to perform reinflection. The Moscow State system employs a multiclass SVM (Bishop, 2006) to solve the selection problem. Overall, this was the best-performing non-neural system. The reason for this may be that the abstract paradigm approach enforces hard constraints between reinflected forms in a way that many of the other non-neural systems do not.

6.4 Performance of Submitted Systems

Relative system performance is described in Table 6, which shows the average rank and per-language accuracy of each system by track and task. The table reflects the fact that some teams submitted more than one system (e.g. LMU-1 & LMU-2 in the table). Full results can be found in the appendix. Table 7 shows that in most cases, competing systems were significantly different (average $p < 0.05$ across 6 unpaired permutation tests for each pair with 5000 permutations per test). The only case in which this did not hold true was in comparing the systems submitted by LMU to one another.

Three teams exploited the bonus resources in some form: LMU, Alberta and Columbia/NYUAD. In general, gains from the bonus resources were modest. Even in Arabic, where the largest benefits were observed, going from track 2 to track 3 on task 1 resulted in an absolute increase in accuracy of only $\approx 3\%$ for LMU's best system.

The neural systems were the clear winner in the shared task. In fact, the gains over classical systems were quite outstanding. The neural systems had two advantages over the competing approaches. First, all these models learned to align and transduce *jointly*. This idea, however, is not intrinsic to neural architectures; it is possible—in fact common—to train finite-state transducers that sum over all possible alignments between the input and output strings (Dreyer et al., 2008; Cotterell et al., 2014).

Second, they all involved massive parameter sharing between the different reinflections. Since the reinflection task entails generalizing from only a few data pairs, this is likely to be a boon. Interestingly, the second BIU-MIT system, which trained a neural model to predict edit operations, consistently ranked behind their first system. This indicates that pre-extracting edit operations, as all systems in the first camp did, is not likely to achieve top-level performance.

Even though the top-ranked neural systems do very well on their own, the other submitted systems may still contain a small amount of complementary information, so that an ensemble over the different approaches has a chance to improve accuracy. We present an upper bound on the possible accuracy of such an ensemble. Table 6 also includes an 'Oracle' that gives the correct answer if *any* of the submitted systems is correct. The average potential ensemble accuracy gain across tasks over the top-ranked system alone is 2.3%. This is the proportion of examples that the top system got wrong, but which some other system got right.

7 Future Directions

Given the success of the submitted reinflection systems in the face of limited data from typologically diverse languages, the future of morphological reinflection must extend in new directions. Further pursuing the line that led us to pose task 3, the problem of morphological reinflection could be expanded by requiring systems to learn with less supervision. Supervised datasets could be smaller or more weakly supervised, forcing systems to rely more on inductive bias or unlabeled data.

One innovation along these lines could be to provide multiple unlabeled source forms and ask for the rest of the paradigm to be produced. In another task, instead of using source and target morphological tags, systems could be asked to induce these from context. Such an extension would necessitate interaction with parsers, and would more closely integrate syntactic and morphological analysis.

Reflecting the traditional linguistic approaches to morphology, another task could allow the use of phonological forms in addition to orthographic forms. While this would necessitate learning a grapheme-to-phoneme mapping, it has the potential to actually simplify the learning task by removing orthographic idiosyncrasies (such as the Spanish '`c`/`qu`' alternation, which is dependent on the backness of the following vowel, but preserves the phoneme /`k`/).

Traditional morphological analyzers, usually

	EHU	BI/M-1	BI/M-2	CU	COL/NYU	HEL	MSU	LMU-1	LMU-2	OSU
UA	90% (10)	—	—	67% (30)	93% (58)	—	79% (28)	100% (60)	100% (30)	—
EHU	—	100% (10)	100% (10)	85% (20)	100% (18)	100% (10)	85% (20)	100% (20)	100% (20)	100% (9)
BI/M-1	—	—	70% (20)	86% (28)	100% (22)	67% (30)	93% (28)	100% (30)	100% (30)	100% (9)
BI/M-2	—	—	—	95% (19)	100% (12)	80% (20)	79% (19)	95% (20)	95% (20)	100% (9)
CU	—	—	—	—	86% (49)	96% (28)	84% (56)	100% (58)	100% (58)	100% (9)
COL/NYU	—	—	—	—	—	95% (22)	96% (47)	100% (80)	100% (50)	100% (8)
HEL	—	—	—	—	—	—	89% (28)	97% (30)	97% (30)	100% (9)
MSU	—	—	—	—	—	—	—	96% (56)	96% (56)	100% (9)
LMU-1	—	—	—	—	—	—	—	—	3% (60)	100% (9)
LMU-2	—	—	—	—	—	—	—	—	—	100% (9)

Table 7: How often each pair of systems had significantly different accuracy under a paired permutation test ($p < 0.05$), as a fraction of the number of times that they competed (on the same language, track and task). The number of such competitions is in parentheses.

implemented as finite state transducers (Beesley and Karttunen, 2003), often return all morphologically plausible analyses if there is ambiguity. Learning to mimic the behavior of a hand-written analyzer in this respect could offer a more challenging task, and one that is useful within unsupervised learning (Dreyer and Eisner, 2011) as well as parsing. Existing wide-coverage morphological analyzers could be leveraged in the design of a more interactive shared task, where hand-coded models or approximate surface rules could serve as informants for grammatical inference algorithms.

The current task design did not explore all potential inflectional complexities in the languages included. For example, cliticization processes were generally not present in the language data. Adding such inflectional elements to the task can potentially make it more realistic in terms of real-world data sparsity in L1 learning scenarios. For example, Finnish noun and adjective inflection is generally modeled as a paradigm of 15 cases in singular and plural, i.e. with 30 slots in total—the shared task data included precisely such paradigms. However, adding all combinations of clitics raises the number of entries in an inflection table to 2,253 (Karlsson, 2008).

Although the languages introduced in this year's shared task were typologically diverse with a range of morphological types (agglutinative, fusional; prefixing, infixing, suffixing, or a mix), we did not cover reduplicative morphology, which is common in Austronesian languages (and elsewhere) but is avoided by traditional computational morphology since it cannot be represented using finite-state transduction. Furthermore, the focus was solely on inflectional data. Another version of the task could call for learning derivational morphology and predicting which derivational forms led to grammatical output (i.e. existing words or neologisms that are not subject to morphological blocking; Poser (1992)). This could be extended to learning the morphology of polysynthetic languages. These languages productively use not only inflection and derivation, which call for the addition of bound morphemes, but also incorporation, which involves combining lexical stems that are often used to form independent words (Mithun, 1984). Such languages combine the need to decompound, generate derivational alternatives, and accurately inflect any resulting words.

8 Conclusion

The SIGMORPHON 2016 Shared Task on Morphological Reinflection significantly expanded the problem of morphological reinflection from a problem of generating complete paradigms from a designated lemma form to generating requested forms based on arbitrary inflected forms, in some cases without a morphological tag identifying the paradigm cell occupied by that form. Furthermore, complete paradigms were not provided in the training data. The submitted systems employed a wide variety of approaches, both neural network-based approaches and extensions of non-neural approaches pursued in previous works such as Durrett and DeNero (2013), Ahlberg et al. (2015), and Nicolai et al. (2015). The superior performance of the neural approaches was likely due to the increased parameter sharing available in those architectures, as well as their ability to discover subtle linguistic features from these relatively large training sets, such as weak or long-distance contextual features that are less likely to appear in hand-engineered feature sets.

References

Roee Aharoni, Yoav Goldberg, and Yonatan Belinkov. 2016. Improving sequence to sequence learning for morphological inflection generation: The BIU-MIT systems for the SIGMORPHON 2016 shared task for morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th EACL*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1024–1029, Denver, CO. Association for Computational Linguistics.

Iñaki Alegria and Izaskun Etxeberria. 2016. EHU at the SIGMORPHON 2016 shared task. A simple proposal: Grapheme-to-phoneme for inflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany, August. Association for Computational Linguistics.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFST: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata, 12th International Conference, CIAA 2007, Prague, Czech Republic, July 16-18, 2007, Revised Selected Papers*, pages 11–23.

Howard I. Aronson. 1990. *Georgian: A Reading Grammar*. Slavica, Columbus, OH.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, CA.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

John J. Camilleri. 2013. A computational grammar and lexicon for Maltese. Master's thesis, Chalmers University of Technology. Gothenburg, Sweden.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning morphology with Morfette. In *LREC*.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2014. Stochastic contextual edit distance and probabilistic fsts. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 625–630, Baltimore, Maryland. Association for Computational Linguistics.

Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics*, 3:433–447.

Markus Dreyer and Jason Eisner. 2009. Graphical models over multiple strings. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 101–110. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of EMNLP 2011*, pages 616–627, Edinburgh. Association for Computational Linguistics.

Markus Dreyer, Jason R. Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *EMNLP*, pages 1080–1089.

Kevin Duh and Katrin Kirchhoff. 2004. Automatic learning of language model structure. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 148–154, Stroudsburg, PA. Association for Computational Linguistics.

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.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *ACL*.

Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic extraction of morphological lexicons from morphologically annotated corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1032–1043.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 676–683, Stroudsburg, PA. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on strings, trees and sequences: Computer science and computational biology*. Cambridge University Press.

Gunnar Ólafur Hansson. 2010. *Consonant Harmony: Long-Distance Interaction in Phonology*. University of California Publications in Linguistics. University of California Press, Berkeley, CA.

Daniel S. Hirschberg. 1975. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341–343.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 372–379.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 905–913.

Katharina Kann and Hinrich Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Fred Karlsson. 2008. *Finnish: An essential grammar*. Routledge.

Aleksandr E. Kibrik. 1998. Archi. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 455–476. Blackwell, Oxford.

David King. 2016. Evaluating sequence alignment for learning inflectional morphology. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany, August. Association for Computational Linguistics.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *arXiv preprint arXiv:1603.04351*.

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 (LREC 2016)*, pages 3121–3126, Paris, France. European Language Resources Association (ELRA).

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Ling Liu and Lingshuang Jack Mao. 2016. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Marianne Mithun. 1984. The evolution of noun incorporation. *Language*, 60(4):847–894, December.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with LEMMING. In *Empirical Methods in Natural Language Processing*.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931.

Garrett Nicolai, Bradley Hauer, Adam St. Arnaud, and Grzegorz Kondrak. 2016. Morphological reinflection via discriminative string transduction. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *10th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP)*, pages 45–49.

Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

William J. Poser. 1992. Blocking of phrasal constructions by lexical items. In Ivan Sag and Anna Szabolcsi, editors, *Lexical Matters*, pages 111–130, Palo Alto, CA. CSLI.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting finite-state transductions with neural context. In *NAACL*.

Keren Rice. 2000. *Morpheme Order and Semantic Scope: Word Formation in the Athapaskan Verb*. Cambridge University Press, Cambridge, UK.

Sunita Sarawagi and William W. Cohen. 2004. Semi-Markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004]*, pages 1185–1192.

Alexey Sorokin. 2016. Using longest common subsequence and character models to predict word forms. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Katsuhito Sudoh, Shinsuke Mori, and Masaaki Nagata. 2013. Noise-aware character alignment for bootstrapping statistical machine transliteration from bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 204–209.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015a. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In Cerstin Mahlow and Michael Piotrowski, editors, *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, pages 72–93. Springer, Berlin.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015b. 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 (ACL-IJCNLP)*, pages 674–680, Beijing. Association for Computational Linguistics.

Dima Taji, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. The Columbia University - New York University Abu Dhabi SIGMORPHON 2016 morphological reinflection shared task submission. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Robert W. Young and William Morgan. 1987. *The Navajo Language: A Grammar and Colloquial Dictionary*. University of New Mexico Press, Albuquerque.

Appendix: Full Results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7
Arabic	LMU-1 95.47 / 0.11 / 95.47	LMU-2 95.47 / 0.11 / 95.47	MSU 83.52 / 0.47 / 86.32	UA 82.05 / 0.36 / 86.79	CU 72.42 / 0.57 / 72.42	EHU 63.1 / 0.65 / 63.1	COL / NYU 57.36 / 1.05 / 57.36
Finnish	LMU-1 96.8 / 0.05 / 96.8	LMU-2 96.8 / 0.05 / 96.8	CU 88.65 / 0.15 / 88.65	UA 88.52 / 0.14 / 92.64	MSU 87.86 / 0.15 / 90.57	EHU 83.72 / 0.24 / 83.72	COL / NYU —
Georgian	LMU-1 98.5 / 0.02 / 98.5	LMU-2 98.5 / 0.02 / 98.5	UA 96.15 / 0.07 / 97.03	MSU 94.8 / 0.1 / 95.98	CU 93.86 / 0.12 / 93.86	COL / NYU 89.49 / 0.22 / 89.49	EHU 82.88 / 0.42 / 82.88
German	LMU-1 95.8 / 0.07 / 95.8	LMU-2 95.8 / 0.07 / 95.8	MSU 94.22 / 0.11 / 95.17	UA 94.06 / 0.1 / 95.82	CU 92.64 / 0.13 / 92.64	EHU 89.71 / 0.16 / 89.71	COL / NYU 86.0 / 0.23 / 86.0
Hungarian	LMU-1 99.3 / 0.01 / 99.3	LMU-2 99.3 / 0.01 / 99.3	COL / NYU 91.8 / 0.18 / 91.8	CU 91.05 / 0.15 / 91.05	MSU 89.82 / 0.16 / 92.69	UA 86.71 / 0.17 / 91.47	EHU 85.0 / 0.24 / 85.0
Maltese	LMU-1 88.99 / 0.27 / 88.99	LMU-2 88.99 / 0.27 / 88.99	EHU 63.09 / 1.07 / 63.09	MSU 52.5 / 1.08 / 57.43	CU 43.49 / 1.45 / 43.49	UA 41.95 / 1.98 / 52.0	COL / NYU 31.03 / 1.48 / 31.03
Navajo	LMU-1 91.48 / 0.17 / 91.48	LMU-2 91.48 / 0.17 / 91.48	UA 60.26 / 1.03 / 68.56	EHU 58.3 / 1.05 / 58.3	MSU 57.42 / 1.48 / 60.74	CU 53.28 / 1.29 / 53.28	COL / NYU 41.27 / 1.81 / 41.27
Russian	LMU-1 91.46 / 0.15 / 91.46	LMU-2 91.46 / 0.15 / 91.46	MSU 89.67 / 0.18 / 91.71	CU 89.13 / 0.18 / 89.13	UA 88.62 / 0.19 / 91.65	EHU 86.22 / 0.24 / 86.22	COL / NYU 60.63 / 0.77 / 60.63
Spanish	LMU-1 98.84 / 0.02 / 98.84	LMU-2 98.84 / 0.02 / 98.84	MSU 98.76 / 0.02 / 98.91	CU 98.28 / 0.03 / 98.28	UA 97.76 / 0.04 / 98.55	EHU 91.16 / 0.19 / 91.16	COL / NYU 76.98 / 0.42 / 76.98
Turkish	LMU-1 98.93 / 0.02 / 98.93	LMU-2 98.93 / 0.02 / 98.93	MSU 92.03 / 0.14 / 92.8	EHU 89.21 / 0.25 / 89.21	CU 87.39 / 0.28 / 87.39	UA 82.25 / 0.33 / 86.42	COL / NYU 76.16 / 0.61 / 76.16

Table 8: Standard track - Task 1 Accuracy / Levenshtein / Reciprocal Rank Results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
Arabic	LMU-1 97.38 / 0.06 / 97.38	LMU-2 97.38 / 0.06 / 97.38	MSU 80.07 / 0.55 / 82.12	UA 71.78 / 0.79 / 76.51	CU 62.74 / 0.79 / 62.74	COL / NYU 61.64 / 1.01 / 61.64
Finnish	LMU-1 97.4 / 0.04 / 97.4	LMU-2 97.4 / 0.04 / 97.4	UA 85.6 / 0.24 / 89.12	MSU 85.18 / 0.2 / 89.2	CU 80.19 / 0.29 / 80.19	COL / NYU —
Georgian	LMU-1 99.14 / 0.01 / 99.14	LMU-2 99.07 / 0.01 / 99.07	MSU 92.39 / 0.17 / 94.14	COL / NYU 91.1 / 0.23 / 91.1	CU 90.87 / 0.18 / 90.87	UA 73.73 / 0.63 / 76.65
German	LMU-1 97.45 / 0.04 / 97.45	LMU-2 97.45 / 0.04 / 97.45	MSU 93.65 / 0.12 / 94.86	UA 91.15 / 0.14 / 93.48	COL / NYU 88.95 / 0.18 / 88.96	CU 88.44 / 0.18 / 88.44
Hungarian	LMU-1 99.67 / 0.01 / 99.67	LMU-2 99.67 / 0.01 / 99.67	COL / NYU 90.54 / 0.23 / 90.54	MSU 89.83 / 0.16 / 92.87	CU 87.49 / 0.21 / 87.49	UA 86.29 / 0.19 / 90.92
Maltese	LMU-1 88.17 / 0.3 / 88.17	LMU-2 88.17 / 0.3 / 88.17	UA 37.5 / 2.14 / 47.1	COL / NYU 29.13 / 1.48 / 29.13	CU 22.54 / 2.12 / 22.54	MSU —
Navajo	LMU-1 96.64 / 0.07 / 96.64	LMU-2 96.64 / 0.07 / 96.64	COL / NYU 84.67 / 0.57 / 84.96	MSU 54.45 / 1.63 / 57.32	UA 50.36 / 1.42 / 60.06	CU 46.13 / 1.62 / 46.2
Russian	LMU-1 91.0 / 0.15 / 91.0	LMU-2 90.32 / 0.16 / 90.32	MSU 88.04 / 0.2 / 90.85	CU 86.71 / 0.21 / 86.71	UA 85.5 / 0.23 / 88.84	COL / NYU 66.0 / 0.69 / 66.0
Spanish	LMU-1 98.74 / 0.02 / 98.74	LMU-2 98.74 / 0.02 / 98.74	MSU 98.21 / 0.03 / 98.5	CU 97.18 / 0.04 / 97.18	UA 96.18 / 0.06 / 97.28	COL / NYU 84.53 / 0.25 / 84.53
Turkish	LMU-1 97.94 / 0.03 / 97.94	LMU-2 97.94 / 0.03 / 97.94	MSU 92.69 / 0.13 / 93.31	COL / NYU 83.75 / 0.54 / 83.75	CU 67.5 / 0.62 / 67.5	UA 62.5 / 1.71 / 66.88

Table 9: Standard track - Task 2 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
Arabic	LMU-1 96.52 / 0.08 / 96.52	LMU-2 96.52 / 0.08 / 96.52	MSU 80.57 / 0.55 / 82.64	COL / NYU 59.4 / 1.07 / 59.4	CU 58.83 / 0.9 / 58.83	UA 23.87 / 3.92 / 25.82
Finnish	LMU-1 96.56 / 0.06 / 96.56	LMU-2 96.56 / 0.06 / 96.56	UA 85.77 / 0.23 / 89.39	MSU 84.34 / 0.22 / 88.41	CU 79.45 / 0.31 / 79.45	COL / NYU 20.3 / 1.64 / 20.3
Georgian	LMU-1 98.87 / 0.02 / 98.87	LMU-2 98.87 / 0.02 / 98.87	UA 94.13 / 0.11 / 95.33	CU 90.43 / 0.19 / 90.43	COL / NYU 90.42 / 0.25 / 90.42	MSU 88.05 / 0.35 / 90.04
German	LMU-1 95.6 / 0.07 / 95.6	LMU-2 95.6 / 0.07 / 95.6	MSU 91.99 / 0.15 / 93.7	UA 91.15 / 0.14 / 93.47	CU 86.59 / 0.21 / 86.59	COL / NYU 78.57 / 0.36 / 78.57
Hungarian	LMU-1 99.5 / 0.01 / 99.5	LMU-2 99.5 / 0.01 / 99.5	MSU 90.41 / 0.16 / 93.12	COL / NYU 89.12 / 0.27 / 89.12	CU 87.04 / 0.23 / 87.04	UA 86.41 / 0.18 / 90.96
Maltese	LMU-1 87.83 / 0.31 / 87.83	LMU-2 87.83 / 0.31 / 87.83	UA 37.5 / 2.14 / 47.1	COL / NYU 28.92 / 1.52 / 28.92	CU 20.58 / 2.22 / 20.58	MSU —
Navajo	LMU-1 96.2 / 0.06 / 96.2	LMU-2 96.2 / 0.06 / 96.2	COL / NYU 84.38 / 0.59 / 84.38	MSU 52.12 / 1.85 / 55.84	UA 48.76 / 1.42 / 59.06	CU 47.3 / 1.61 / 47.3
Russian	LMU-1 89.91 / 0.17 / 89.91	LMU-2 89.91 / 0.17 / 89.91	MSU 86.59 / 0.23 / 89.75	UA 85.48 / 0.23 / 88.85	CU 85.34 / 0.23 / 85.34	COL / NYU 62.07 / 0.73 / 62.07
Spanish	LMU-1 97.96 / 0.03 / 97.96	LMU-2 97.96 / 0.03 / 97.96	MSU 97.3 / 0.04 / 97.85	UA 96.27 / 0.06 / 97.34	CU 96.26 / 0.06 / 96.26	COL / NYU 78.55 / 0.41 / 78.55
Turkish	LMU-1 99.31 / 0.01 / 99.31	LMU-2 99.31 / 0.01 / 99.31	MSU 92.44 / 0.14 / 93.16	COL / NYU 84.38 / 0.32 / 84.38	CU 65.63 / 0.65 / 65.63	UA 63.13 / 1.66 / 67.29

Table 10: Standard track - Task 3 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
Arabic	LMU-1 95.47 / 0.11 / 95.47	LMU-2 95.47 / 0.11 / 95.47	BIU / MIT-1 93.34 / 0.16 / 94.74	BIU / MIT-2 89.96 / 0.19 / 89.96	HEL 89.52 / 0.23 / 89.52	MSU 83.52 / 0.47 / 86.32	CU 72.42 / 0.57 / 72.42	EHU 64.68 / 0.62 / 64.68	COL / NYU 57.36 / 1.05 / 57.36	OSU 0.82 / 4.59 / 0.82
Finnish	LMU-1 96.8 / 0.05 / 96.8	LMU-2 96.8 / 0.05 / 96.8	HEL 95.14 / 0.06 / 95.14	BIU / MIT-1 93.81 / 0.09 / 95.43	BIU / MIT-2 92.58 / 0.12 / 92.58	CU 88.65 / 0.15 / 88.65	MSU 87.86 / 0.15 / 90.57	EHU 83.72 / 0.24 / 83.72	OSU 2.01 / 3.48 / 2.01	COL / NYU —
Georgian	LMU-1 98.5 / 0.02 / 98.5	LMU-2 98.5 / 0.02 / 98.5	BIU / MIT-1 97.55 / 0.03 / 98.17	HEL 97.02 / 0.05 / 97.02	BIU / MIT-2 96.54 / 0.05 / 96.54	MSU 94.8 / 0.1 / 95.98	CU 93.86 / 0.12 / 93.86	COL / NYU 89.49 / 0.22 / 89.49	EHU 83.11 / 0.42 / 83.11	OSU 10.75 / 2.08 / 10.75
German	LMU-1 95.8 / 0.07 / 95.8	LMU-2 95.8 / 0.07 / 95.8	BIU / MIT-1 95.11 / 0.09 / 96.12	BIU / MIT-2 94.87 / 0.09 / 94.87	HEL 94.4 / 0.09 / 94.4	MSU 94.22 / 0.11 / 95.17	CU 92.64 / 0.13 / 92.64	EHU 89.86 / 0.16 / 89.86	COL / NYU 86.0 / 0.23 / 86.0	OSU 11.01 / 1.99 / 11.01
Hungarian	LMU-1 99.3 / 0.01 / 99.3	LMU-2 99.3 / 0.01 / 99.3	HEL 98.38 / 0.02 / 98.38	BIU / MIT-1 98.33 / 0.02 / 98.79	BIU / MIT-2 97.59 / 0.04 / 97.59	COL / NYU 91.8 / 0.18 / 91.8	CU 91.05 / 0.15 / 91.05	MSU 89.82 / 0.16 / 92.69	EHU 85.39 / 0.23 / 85.39	OSU 1.62 / 3.3 / 1.62
Maltese	LMU-1 88.99 / 0.27 / 88.99	LMU-2 88.99 / 0.27 / 88.99	HEL 86.16 / 0.3 / 86.16	BIU / MIT-2 84.78 / 0.41 / 84.78	BIU / MIT-1 82.4 / 0.41 / 88.76	EHU 64.8 / 0.98 / 64.8	MSU 52.5 / 1.08 / 57.43	CU 43.49 / 1.45 / 43.49	COL / NYU 31.03 / 1.48 / 31.03	OSU —
Navajo	LMU-1 91.48 / 0.17 / 91.48	LMU-2 91.48 / 0.17 / 91.48	BIU / MIT-2 88.43 / 0.22 / 88.43	HEL 82.1 / 0.33 / 82.1	BIU / MIT-1 80.13 / 0.4 / 83.54	MSU 57.42 / 1.48 / 60.74	EHU 56.33 / 1.11 / 56.33	CU 53.28 / 1.29 / 53.28	COL / NYU 41.27 / 1.81 / 41.27	OSU 2.18 / 4.6 / 2.18
Russian	LMU-1 91.46 / 0.15 / 91.46	LMU-2 91.46 / 0.15 / 91.46	BIU / MIT-2 90.62 / 0.16 / 90.62	HEL 89.94 / 0.17 / 89.94	BIU / MIT-1 89.73 / 0.18 / 91.96	MSU 89.67 / 0.18 / 91.71	CU 89.13 / 0.18 / 89.13	EHU 86.58 / 0.23 / 86.58	COL / NYU 60.63 / 0.77 / 60.63	OSU 5.49 / 1.92 / 5.49
Spanish	LMU-1 98.84 / 0.02 / 98.84	LMU-2 98.84 / 0.02 / 98.84	MSU 98.76 / 0.02 / 98.91	BIU / MIT-2 98.41 / 0.02 / 98.41	HEL 98.35 / 0.03 / 98.35	BIU / MIT-1 98.33 / 0.03 / 98.7	CU 98.28 / 0.03 / 98.28	EHU 91.35 / 0.23 / 91.35	COL / NYU 76.98 / 0.42 / 76.98	OSU 8.52 / 2.24 / 8.52
Turkish	LMU-1 98.93 / 0.02 / 98.93	LMU-2 98.93 / 0.02 / 98.93	HEL 97.93 / 0.03 / 97.93	BIU / MIT-1 97.74 / 0.03 / 98.36	BIU / MIT-2 96.17 / 0.07 / 96.17	MSU 92.03 / 0.14 / 92.8	EHU 90.84 / 0.21 / 90.84	CU 87.39 / 0.28 / 87.39	COL / NYU 76.16 / 0.61 / 76.16	OSU 2.01 / 4.59 / 2.01

Table 11: Restricted track - Task 1 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
Arabic	LMU-1 91.09 / 0.2 / 91.09	LMU-2 91.09 / 0.2 / 91.09	BIU / MIT-2 73.95 / 0.53 / 73.95	BIU / MIT-1 70.26 / 0.6 / 75.62	HEL 69.53 / 0.67 / 69.53	MSU 52.88 / 1.23 / 61.4	CU 32.86 / 1.96 / 32.86	COL / NYU —
Finnish	LMU-1 96.81 / 0.04 / 96.81	LMU-2 96.81 / 0.04 / 96.81	HEL 88.42 / 0.15 / 88.42	BIU / MIT-2 80.91 / 0.29 / 80.91	MSU 77.53 / 0.3 / 84.92	BIU / MIT-1 74.44 / 0.38 / 80.11	CU 66.24 / 0.52 / 66.24	COL / NYU —
Georgian	LMU-1 98.5 / 0.02 / 98.5	LMU-2 98.37 / 0.02 / 98.37	HEL 92.84 / 0.15 / 92.84	BIU / MIT-1 92.65 / 0.14 / 93.75	BIU / MIT-2 92.27 / 0.15 / 92.27	COL / NYU 90.21 / 0.26 / 90.21	CU 89.95 / 0.28 / 89.95	MSU 86.22 / 0.31 / 89.86
German	LMU-1 96.22 / 0.06 / 96.22	LMU-2 96.22 / 0.06 / 96.22	BIU / MIT-2 92.66 / 0.14 / 92.66	MSU 92.39 / 0.14 / 94.4	HEL 91.73 / 0.15 / 91.73	BIU / MIT-1 91.67 / 0.13 / 93.56	CU 89.37 / 0.17 / 89.37	COL / NYU —
Hungarian	LMU-1 99.42 / 0.01 / 99.42	LMU-2 99.42 / 0.01 / 99.42	HEL 96.25 / 0.05 / 96.25	BIU / MIT-1 92.33 / 0.11 / 94.43	BIU / MIT-2 91.16 / 0.14 / 91.16	MSU 90.45 / 0.16 / 93.31	CU 83.58 / 0.3 / 83.58	COL / NYU —
Maltese	LMU-1 86.88 / 0.3 / 86.88	LMU-2 86.88 / 0.3 / 86.88	HEL 73.17 / 0.54 / 73.17	BIU / MIT-2 50.13 / 1.24 / 50.13	BIU / MIT-1 41.92 / 1.45 / 50.79	COL / NYU —	MSU —	CU —
Navajo	LMU-1 97.81 / 0.05 / 97.81	LMU-2 97.81 / 0.05 / 97.81	COL / NYU 80.88 / 0.76 / 81.17	HEL 77.37 / 0.44 / 77.37	MSU 69.93 / 0.95 / 77.87	BIU / MIT-2 67.88 / 0.62 / 67.88	BIU / MIT-1 56.5 / 0.98 / 62.54	CU 53.58 / 1.31 / 53.65
Russian	LMU-1 90.11 / 0.16 / 90.11	LMU-2 89.43 / 0.17 / 89.43	HEL 86.6 / 0.21 / 86.6	BIU / MIT-2 85.81 / 0.22 / 85.81	MSU 85.67 / 0.24 / 89.79	BIU / MIT-1 83.36 / 0.26 / 86.85	CU 83.11 / 0.28 / 83.11	COL / NYU 60.92 / 0.75 / 60.92
Spanish	LMU-1 98.45 / 0.02 / 98.45	LMU-2 98.45 / 0.02 / 98.45	HEL 95.35 / 0.07 / 95.35	BIU / MIT-2 94.26 / 0.08 / 94.26	BIU / MIT-1 92.21 / 0.11 / 94.05	CU 88.65 / 0.2 / 88.65	MSU 88.38 / 0.16 / 92.85	COL / NYU —
Turkish	LMU-1 98.38 / 0.02 / 98.38	LMU-2 98.38 / 0.02 / 98.38	HEL 91.69 / 0.12 / 91.69	BIU / MIT-2 83.88 / 0.34 / 83.88	BIU / MIT-1 81.69 / 0.28 / 86.48	MSU 73.69 / 0.62 / 81.77	CU 41.69 / 1.79 / 41.69	COL / NYU —

Table 12: Restricted track - Task 2 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7
Arabic	LMU-1 82.8 / 0.4 / 82.8	LMU-2 82.8 / 0.4 / 82.8	HEL 70.43 / 0.67 / 70.43	BIU / MIT-1 69.05 / 0.65 / 74.76	COL / NYU 47.46 / 1.44 / 47.46	CU 29.99 / 2.04 / 29.99	MSU 26.23 / 2.27 / 28.13
Finnish	LMU-1 93.18 / 0.11 / 93.18	LMU-2 93.18 / 0.11 / 93.18	HEL 87.55 / 0.17 / 87.55	BIU / MIT-1 72.99 / 0.41 / 78.93	CU 65.28 / 0.54 / 65.28	MSU 54.98 / 0.9 / 59.06	COL / NYU 20.3 / 1.64 / 20.3
Georgian	LMU-1 96.21 / 0.06 / 96.21	LMU-2 96.21 / 0.06 / 96.21	BIU / MIT-1 92.08 / 0.15 / 93.38	HEL 91.85 / 0.17 / 91.85	CU 89.27 / 0.3 / 89.27	COL / NYU 87.79 / 0.29 / 87.79	MSU 79.94 / 0.6 / 82.02
German	LMU-1 92.41 / 0.13 / 92.41	LMU-2 92.41 / 0.13 / 92.41	BIU / MIT-1 89.58 / 0.17 / 91.67	HEL 89.14 / 0.2 / 89.14	MSU 89.07 / 0.2 / 91.84	CU 85.45 / 0.23 / 85.45	COL / NYU 75.39 / 0.42 / 75.39
Hungarian	LMU-1 98.37 / 0.03 / 98.37	LMU-2 98.37 / 0.03 / 98.37	HEL 96.46 / 0.06 / 96.46	BIU / MIT-1 91.91 / 0.11 / 93.95	CU 82.41 / 0.32 / 82.41	MSU 80.45 / 0.45 / 81.91	COL / NYU 70.49 / 0.74 / 70.49
Maltese	LMU-1 84.25 / 0.36 / 84.25	LMU-2 84.25 / 0.36 / 84.25	HEL 75.54 / 0.5 / 75.54	BIU / MIT-1 40.79 / 1.48 / 49.49	COL / NYU 24.21 / 1.89 / 24.21	MSU —	CU —
Navajo	LMU-1 83.5 / 0.38 / 83.5	LMU-2 83.5 / 0.38 / 83.5	HEL 83.21 / 0.36 / 83.21	BIU / MIT-1 52.85 / 1.07 / 59.07	CU 50.07 / 1.51 / 50.07	MSU 35.33 / 2.31 / 37.64	COL / NYU 29.34 / 3.33 / 29.34
Russian	LMU-1 87.13 / 0.21 / 87.13	LMU-2 87.13 / 0.21 / 87.13	HEL 84.59 / 0.24 / 84.59	BIU / MIT-1 82.81 / 0.27 / 86.47	CU 80.77 / 0.34 / 80.77	MSU 79.59 / 0.39 / 85.2	COL / NYU 52.56 / 0.93 / 52.56
Spanish	LMU-1 96.69 / 0.05 / 96.69	LMU-2 96.69 / 0.05 / 96.69	HEL 94.85 / 0.08 / 94.85	BIU / MIT-1 92.14 / 0.12 / 94.08	CU 87.92 / 0.23 / 87.92	MSU 82.93 / 0.36 / 84.98	COL / NYU 68.07 / 0.58 / 68.07
Turkish	LMU-1 95.0 / 0.08 / 95.0	LMU-2 95.0 / 0.08 / 95.0	HEL 91.25 / 0.14 / 91.25	BIU / MIT-1 79.69 / 0.32 / 85.15	COL / NYU 59.94 / 0.9 / 59.94	CU 40.06 / 1.85 / 40.06	MSU 30.88 / 2.6 / 31.82

Table 13: Restricted track - Task 3 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3
Arabic	LMU-1 98.25 / 0.04 / 98.25	COL / NYU 69.31 / 0.68 / 69.31	UA 53.12 / 0.81 / 71.84
Finnish	LMU-1 97.3 / 0.04 / 97.3	UA 88.74 / 0.14 / 92.84	COL / NYU 15.36 / 1.92 / 15.36
Georgian	LMU-1 99.2 / 0.01 / 99.2	UA 96.3 / 0.07 / 97.14	COL / NYU 94.01 / 0.14 / 94.01
German	LMU-1 97.38 / 0.05 / 97.38	UA 93.8 / 0.11 / 95.7	COL / NYU 88.38 / 0.19 / 88.38
Hungarian	LMU-1 99.69 / 0.01 / 99.69	COL / NYU 94.12 / 0.13 / 94.12	UA 89.61 / 0.13 / 93.01
Maltese	LMU-1 88.53 / 0.29 / 88.53	UA 42.45 / 1.97 / 52.28	COL / NYU 35.53 / 1.16 / 35.53
Navajo	LMU-1 98.03 / 0.04 / 98.03	COL / NYU 87.55 / 0.32 / 87.55	UA 60.26 / 1.03 / 68.6
Russian	LMU-1 92.15 / 0.13 / 92.15	UA 89.67 / 0.18 / 92.4	COL / NYU 67.31 / 0.62 / 67.31
Spanish	LMU-1 99.05 / 0.01 / 99.05	UA 98.03 / 0.03 / 98.72	COL / NYU 84.44 / 0.28 / 84.44
Turkish	LMU-1 97.49 / 0.04 / 97.49	UA 87.52 / 0.23 / 90.11	COL / NYU 87.01 / 0.3 / 87.01

Table 14: Bonus track - Task 1 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3
Arabic	LMU-1 97.38 / 0.06 / 97.38	COL / NYU 61.64 / 1.01 / 61.64	UA 44.06 / 1.45 / 62.1
Finnish	LMU-1 97.4 / 0.04 / 97.4	UA 85.65 / 0.25 / 89.15	COL / NYU 37.08 / 1.23 / 37.08
Georgian	LMU-1 99.14 / 0.01 / 99.14	COL / NYU 91.1 / 0.23 / 91.1	UA 74.5 / 0.62 / 77.2
German	LMU-1 97.45 / 0.04 / 97.45	UA 91.58 / 0.14 / 93.77	COL / NYU 88.95 / 0.18 / 88.96
Hungarian	LMU-1 99.67 / 0.01 / 99.67	COL / NYU 90.54 / 0.23 / 90.54	UA 88.79 / 0.14 / 92.31
Maltese	LMU-1 88.17 / 0.3 / 88.17	UA 37.83 / 2.12 / 47.3	COL / NYU 29.13 / 1.48 / 29.13
Navajo	LMU-1 96.64 / 0.07 / 96.64	COL / NYU 84.67 / 0.57 / 84.96	UA 50.8 / 1.41 / 60.12
Russian	LMU-1 91.0 / 0.15 / 91.0	UA 86.6 / 0.22 / 89.77	COL / NYU 66.0 / 0.69 / 66.0
Spanish	LMU-1 98.74 / 0.02 / 98.74	UA 96.52 / 0.06 / 97.54	COL / NYU 84.53 / 0.25 / 84.53
Turkish	LMU-1 97.94 / 0.03 / 97.94	COL / NYU 83.75 / 0.54 / 83.75	UA 59.25 / 1.93 / 65.21

Table 15: Bonus track - Task 2 Accuracy / Levenshtein / Reciprocal Rank results

	Rank 1	Rank 2	Rank 3
Arabic	LMU-1 96.52 / 0.08 / 96.52	COL / NYU 59.4 / 1.07 / 59.4	UA 10.45 / 4.41 / 18.83
Finnish	LMU-1 96.56 / 0.06 / 96.56	UA 85.89 / 0.23 / 89.45	COL / NYU 20.3 / 1.64 / 20.3
Georgian	LMU-1 98.87 / 0.02 / 98.87	UA 94.36 / 0.11 / 95.5	COL / NYU 90.42 / 0.25 / 90.42
German	LMU-1 95.6 / 0.07 / 95.6	UA 91.56 / 0.14 / 93.76	COL / NYU 78.57 / 0.36 / 78.57
Hungarian	LMU-1 99.5 / 0.01 / 99.5	COL / NYU 89.12 / 0.27 / 89.12	UA 88.91 / 0.14 / 92.35
Maltese	LMU-1 87.83 / 0.31 / 87.83	UA 37.83 / 2.12 / 47.3	COL / NYU 28.92 / 1.52 / 28.92
Navajo	LMU-1 96.2 / 0.06 / 96.2	COL / NYU 84.38 / 0.59 / 84.38	UA 49.05 / 1.42 / 58.98
Russian	LMU-1 89.91 / 0.17 / 89.91	UA 86.62 / 0.22 / 89.79	COL / NYU 62.07 / 0.73 / 62.07
Spanish	LMU-1 97.96 / 0.03 / 97.96	UA 96.58 / 0.05 / 97.59	COL / NYU 78.55 / 0.41 / 78.55
Turkish	LMU-1 99.31 / 0.01 / 99.31	COL / NYU 84.38 / 0.32 / 84.38	UA 59.19 / 1.91 / 65.44

Table 16: Bonus track - Task 3 Accuracy / Levenshtein / Reciprocal Rank results